# Calibration estimation in dual frame surveys

M. Giovanna Ranalli* Antonio Arcos† María del Mar Rueda‡ Annalisa Teodoro§

April 8, 2014

**Abstract**

Survey statisticians make use of the available auxiliary information to improve estimates. One important example is given by calibration estimation, that seeks for new weights that are close (in some sense) to the basic design weights and that, at the same time, match benchmark constraints on available auxiliary information. Recently, multiple frame surveys have gained much attention and became largely used by statistical agencies and private organizations to decrease sampling costs or to reduce frame undercoverage errors that could occur with the use of only a single sampling frame. Much attention has been devoted to the introduction of different ways of combining estimates coming from the different frames. We will extend the calibration paradigm, developed so far for one frame surveys, to the estimation of the total of a variable of interest in dual frame surveys as a general tool to include auxiliary information, also available at different levels. In fact, calibration allows us to handle different types of auxiliary information and can be shown to encompass as a special cases some of the methods already proposed in the literature. The theoretical properties of the proposed class of estimators are derived and discussed, a set of simulation studies is conducted to compare the efficiency of the procedure in presence of different sets of auxiliary variables. Finally, the proposed methodology is applied to data from the Barometer of Culture of Andalusia survey.

**Keywords**: Auxiliary information, Kullback-Leibler distance, Raking ratio, Regression estimation, Survey Methodology, Unequal probability sampling.

## 1 Introduction

A main aim of survey statisticians is to obtain more accurate estimates, without increasing survey costs. Two popular tools to achieve this goal are $(i)$ the use of more than one population frame to select independent samples and $(ii)$ the use of auxiliary information either at the design or at the estimation stage. The use of more than one list of population units is important because a common practical problem in conducting sample surveys is that frames may be incomplete or out of date, so that resulting estimates may be seriously biased. Multiple frame surveys are useful when no single

*Department of Political Sciences, Università degli Studi di Perugia, Italy, giovanna.ranalli@stat.unipg.it
†Department of Statistics and Operational Research, Universidad de Granada, Spain
‡Department of Statistics and Operational Research, Universidad de Granada, Spain
§Department of Economics, Finance and Statistics, Università degli Studi di Perugia, Italy

frame covers the whole target population but the union of several available frames does, or when information about a subgroup of particular interest comes only from an incomplete frame. They also have other advantages. In fact, Hartley (1962) introduces dual frame surveys as a cost-saving device, showing that they can often achieve the same precision as a single-frame survey at a much reduced cost. Kalton and Anderson (1986) suggest using two frames for sampling rare populations where even greater efficiencies can be obtained. Several estimators of the population total and mean have been proposed in the literature in dual frame surveys, usually classified, according to the level of frame information needed, as *dual-frame* and *single-frame* estimators.

On the other hand, the growing availability of information coming from census data, administrative registers and previous surveys provide a wide range of variables, concerning the population of interest, that are eligible to be employed as auxiliary information to increase efficiency in the estimation procedure. In this scenario, a very relevant example is given by *calibration estimation* that adjusts basic design weights to account for auxiliary information and meet benchmark constraints on auxiliary variables population statistics (Deville and Särndal, 1992). Särndal (2007) provides an overview on developments in calibration estimation. In this paper, we will show how to extend calibration estimation to handle estimation from two frame surveys and how different types of auxiliary information can be easily integrated in the calibration process as benchmark constraints. Moreover, depending on the information available at the design stage, we show how to build calibration estimators under both the dual and the single frame approach. We will show that the proposed class of calibration estimators encompasses as particular cases some of the estimators already proposed in the literature. To show evidence of such connections, we will follow the minimum distance approach for calibration estimation, although using the instrumental variable approach is of course possible.

The paper is organized as follows. In Section 2 notation is introduced and those methods proposed in the literature to handle dual frame estimation are briefly reviewed. Then Section 3 illustrates the proposed class of calibration estimators by first dealing with the dual-frame approach

and then moving, in Section 4, to the single-frame approach. The general form is provided and particular cases are derived according to relevant examples of auxiliary information. The theoretical properties of the proposed estimators are investigated in an asymptotic framework adapted from that of Isaki and Fuller (1982). In addition, analytic and Jackknife variance estimators are proposed. Then, Section 6 reports the results of an extensive simulation study run on a set of synthetic finite populations in which the performance of the proposed class of estimators is investigated for finite size samples. Section 7 shows the application of the proposed estimation technique to data from the Barometer of Culture of Andalusia survey. Section 8 provides some conclusions and directions for future research.

## 2 Estimation in dual frame surveys

Consider a finite set of $N$ population units identified by the integers, $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$, and let $A$ and $B$ be two sampling-frames, both can be incomplete, but it is assumed that together they cover the entire finite population. Let $\mathcal{A}$ be the set of population units in frame $A$ and $\mathcal{B}$ the set of population units in frame $B$. The population of interest, $\mathcal{U}$, may be divided into three mutually exclusive domains, $a = \mathcal{A} \cap \mathcal{B}^c, b = \mathcal{A}^c \cap \mathcal{B}$ and $ab = \mathcal{A} \cap \mathcal{B}$. Because the population units in the overlap domain $ab$ can be sampled in either survey or both surveys, it is convenient to create a duplicate domain $ba = \mathcal{B} \cap \mathcal{A}$, which is identical to $ab = \mathcal{A} \cap \mathcal{B}$, to denote the domain in the overlapping area coming from frame $B$. Let $N$, $N_A$, $N_B$, $N_a$, $N_b$, $N_{ab}$, $N_{ba}$ be the number of population units in $\mathcal{U}$, $\mathcal{A}$, $\mathcal{B}$, $a$, $b$, $ab$, $ba$, respectively. It follows that $N_A = N_a + N_{ab}$, $N_B = N_b + N_{ba}$ and $N = N_a + N_b + N_{ab} = N_a + N_b + N_{ba}$.

Let $y$ be a variable of interest in the population and $y_k$ its value on unit $k$, for $k = 1, \ldots, N$. The entire set of population $y$ values is our finite population $\mathcal{F}$. The objective is to estimate the finite population total $Y = \sum_{k=1}^{N} y_k$ of $y$, that can be written as

$$Y = Y_a + \eta Y_{ab} + (1 - \eta) Y_{ba} + Y_b, \tag{1}$$

where $0 \leq \eta \leq 1$, and $Y_a = \sum_{k \in a} y_k$, $Y_{ab} = \sum_{k \in ab} y_k$, $Y_{ba} = \sum_{k \in ba} y_k$ and $Y_b = \sum_{k \in b} y_k$. Two

probability samples $s_A$ and $s_B$ are drawn independently from frame $A$ and frame $B$ of sizes $n_A$ and $n_B$, respectively. Each design induces first-order inclusion probabilities $\pi_{Ak}$ and $\pi_{Bk}$, respectively, and sampling weights $d_{Ak} = 1/\pi_{Ak}$ and $d_{Bk} = 1/\pi_{Bk}$. Units in $s_A$ can be divided as $s_A = s_a \cup s_{ab}$, where $s_a = s_A \cap a$ and $s_{ab} = s_A \cap (ab)$. Similarly, $s_B = s_b \cup s_{ba}$, where $s_b = s_B \cap b$ and $s_{ba} = s_B \cap (ba)$. Note that $s_{ab}$ and $s_{ba}$ are both from the same domain $ab$, but $s_{ab}$ is part of the frame $A$ sample and $s_{ba}$ is part of the frame $B$ sample. In this way, we have a sort of "poststratified" sample $s = s_a \cup s_{ab} \cup s_{ba} \cup s_b$ with "poststratum" sample sizes $n_a$, $n_{ab}$, $n_{ba}$ and $n_b$. Note that $n_A = n_a + n_{ab}$ and $n_B = n_b + n_{ba}$ (see Rao and Wu, 2010).

The Hartley (1962) estimator of $Y$ is given by

$$\hat{Y}_H(\eta) = \hat{Y}_a + \eta\hat{Y}_{ab} + (1-\eta)\hat{Y}_{ba} + \hat{Y}_b, \tag{2}$$

where $\hat{Y}_a = \sum_{k \in s_a} d_{Ak} y_k$ is the Horvitz-Thompson estimator for the total of domain $a$ and similarly for the other domains. If we let

$$d_k^\circ = \begin{cases} d_{Ak} & \text{if } k \in s_a \\ \eta d_{Ak} & \text{if } k \in s_{ab} \\ (1-\eta) d_{Bk} & \text{if } k \in s_{ba} \\ d_{Bk} & \text{if } k \in s_b \end{cases}$$

then $\hat{Y}_H(\eta) = \sum_{k \in s} d_k^\circ y_k$. In the following, we will drop $\eta$ for ease of notation. Since each domain is estimated by its Horvitz-Thompson estimator, $\hat{Y}_H$ is an unbiased estimator of $Y$ for a given $\eta$. Since frames $A$ and $B$ are sampled independently, the variance of $\hat{Y}_H$ is given by

$$V(\hat{Y}_H) = V(\hat{Y}_a + \eta\hat{Y}_{ab}) + V((1-\eta)\hat{Y}_{ba} + \hat{Y}_b), \tag{3}$$

where the first component of the right hand side is computed under $p_A(\cdot)$ (the sampling design in frame $A$) and the second one under $p_B(\cdot)$, and both are always understood conditional on the finite population $\mathcal{F}$.

Choice of a value for $\eta$ has attracted much attention in literature; the value of $\eta$ that minimizes the variance in (3) depends on unknown population variances and covariances and, when estimated from the data, it depends on the values of the variable of interest. This implies a need to recompute

weights for every variable of interest $y$, which will be inconvenient in practice for statistical agencies conducting surveys with numerous variables and lead to inconsistencies in the estimates (see Lohr, 2009, for a review).

The estimator developed by Fuller and Burmeister (1972, FB) incorporates information regarding the estimation of $N_{ab}$ to improve over $\hat{Y}_H$, but has the drawback of not being a linear combination of $y$ values, unless using simple random sampling. Skinner and Rao (1996) propose a modification of the estimator proposed by Fuller and Burmeister (1972) for simple random sampling to handle complex designs. They introduce a pseudo maximum likelihood (PML) estimator that does not achieve optimality like the FB estimator, but it can be written as a linear combination of the observations and the same set of weights can be used for all variables of interest.

Recently, Rao and Wu (2010) extend the Pseudo-Empirical-Likelihood approach (PEL) proposed by Wu and Rao (2006) from one-frame surveys to dual-frame surveys following a stratification approach. They consider estimation of the population mean of $y$,

$$\bar{Y} = W_a\bar{Y}_a + W_{ab}(\eta)\bar{Y}_{ab} + W_{ba}(\eta)\bar{Y}_{ba} + W_b\bar{Y}_b,$$

where $W_a = N_a/N$, $W_{ab}(\eta) = \eta N_{ab}/N$, $W_{ba}(\eta) = (1-\eta)N_{ab}/N$ and $W_b = N_b/N$, $\bar{Y}_{ab} = \bar{Y}_{ba}$, and again $\eta \in (0,1)$ is a fixed constant to be specified. The PEL function takes the following expression:

$$\begin{aligned} l_D(p_{ak}, p_{abk}, p_{bak}, p_{bk}) \;=\; & n\Big[W_a \sum_{k\in s_a} \tilde{d}_{ak}\log(p_{ak}) + W_{ab}(\eta)\sum_{k\in s_{ab}} \tilde{d}_{abk}\log(p_{abk}) + \\ & + W_{ba}(\eta)\sum_{k\in s_{ba}} \tilde{d}_{bak}\log(p_{bak}) + W_b\sum_{k\in s_b}\tilde{d}_{bk}\log(p_{bk})\Big], \end{aligned} \tag{4}$$

for all $k \in s$, where $n = n_A + n_B$, $\tilde{d}_{a_k} = d_{Ak}/\sum_{k\in s_a} d_{Ak}$, $\tilde{d}_{abk} = d_{Ak}/\sum_{k\in s_{ab}} d_{Ak}$, $\tilde{d}_{bk} = d_{Bk}/\sum_{k\in s_b} d_{Bk}$ and $\tilde{d}_{bak} = d_{Bk}/\sum_{k\in s_{ba}} d_{Bk}$. The four sets of probability measures in (4) are found by maximizing the PEL function under the following normalizing constraints

$$\sum_{k\in s_a} p_{ak} = 1, \qquad \sum_{k\in s_{ab}} p_{abk} = 1, \quad \sum_{k\in s_{ba}} p_{bak} = 1, \qquad \sum_{k\in s_b} p_{bk} = 1,$$

and the constraint induced by the common domain mean $\bar{Y}_{ab} = \bar{Y}_{ba}$

$$\sum_{k\in s_{ab}} p_{abk}y_k = \sum_{k\in s_{ba}} p_{bak}y_k. \tag{5}$$

The maximum PEL estimator of $\bar{Y}$ is then computed as

$$\hat{\bar{Y}}_P = W_a\hat{\bar{Y}}_a + W_{ab}(\eta)\hat{\bar{Y}}_{ab} + W_{ba}(\eta)\hat{\bar{Y}}_{ba} + W_b\hat{\bar{Y}}_b, \tag{6}$$

where $\hat{\bar{Y}}_a = \sum_{k\in s_a}\hat{p}_{ak}y_k$, $\hat{\bar{Y}}_b = \sum_{k\in s_b}\hat{p}_{bk}y_k$ and $\hat{\bar{Y}}_{ab} = \sum_{k\in s_{ab}}\hat{p}_{abk}y_k = \hat{\bar{Y}}_{ba}$ because of constraint (5). Situations in which population domain sizes are not known are sketched and the choice of $\eta$ is also discussed.

When inclusion probabilities in domain $ab$ are known for both frames, and not just for the frame from which the unit was selected, *single-frame* methods can be used that combine the observations into a single dataset and adjust the weights in the intersection domain for multiplicity. In particular, observations from frame $A$ and frame $B$ are combined and the two samples drawn independently from $A$ and $B$ are considered as a single stratified sample over the three domains $a$, $b$ and $ab$. To adjust for multiplicity, the weights are defined as follows for all units in frame $A$ and in frame $B$,

$$d_k^\star = \begin{cases} d_{Ak} & \text{if } k \in s_a \\ (1/d_{Ak} + 1/d_{Bk})^{-1} & \text{if } k \in s_{ab}\cup s_{ba} \\ d_{Bk} & \text{if } k \in s_b \end{cases}.$$

Note that units in the overlap domain, which are expected to be selected a number of times given by $1/d_{Ak} + 1/d_{Bk}$ have equal weights in frame $A$ and in frame $B$. The estimator proposed by Kalton and Anderson (1986) is essentially an Horvitz-Thompson estimator for which

$$\hat{Y}_S = \sum_{k\in s} d_k^\star y_k. \tag{7}$$

Its variance is given by $V(\hat{Y}_S) = V(\sum_{k\in s_A} d_k^\star y_k) + V(\sum_{k\in s_B} d_k^\star y_k)$, where the first component of the right hand side is computed under $p_A(\cdot)$ and the second one under $p_B(\cdot)$. If $N_A$ and $N_B$ were known, the single-frame estimator $\hat{Y}_S$ could be adjusted using raking ratio estimation (Bankier, 1986; Skinner, 1991).

In the following section calibration estimation for dual frame surveys is introduced. We will first consider dual-frame methods and, then, to encompass situations in which auxiliary information is also in the form of inclusion probabilities for all units in both frames from both sampling design, single-frame methods will be considered as well (Section 4).

## 3 Calibration estimation: dual-frame methods

In this section, we will show how to extend calibration estimation, as discussed in one frame surveys by Deville and Särndal (1992), to handle estimation from two frame surveys and how different types of auxiliary information can be easily integrated in the calibration process as benchmark constraints. Now, let $\boldsymbol{x}_k = (x_{1k}, \ldots, x_{pk})$ be the value taken on unit $k$ by a vector of auxiliary variables $\boldsymbol{x}$ of which we assume to know the population total $\boldsymbol{t}_x = \sum_{k=1}^{N} \boldsymbol{x}_k$. This vector of totals may pertain only $\mathcal{A}$, only $\mathcal{B}$, the entire population $\mathcal{U}$, or a combination of the three. We will first look at a general formulation of the problem, and then provide (relevant) examples of auxiliary vectors $\boldsymbol{x}$. Using the calibration paradigm, we wish to modify, as little as possible, basic Hartley weights $d_k^\circ$ to obtain new weights $w_k^\circ$, for $k \in s$ to account for auxiliary information and derive a more accurate estimation of the total $Y$. A general dual-frame calibration estimator can be defined as

$$\hat{Y}_{\text{CAL}} = \sum_{k \in s} w_k^\circ y_k \tag{8}$$

where $w_k^\circ$ is such that

$$\min \sum_{k \in s} G(w_k^\circ, d_k^\circ) \qquad \text{s.t.} \qquad \sum_{k \in s} w_k^\circ \boldsymbol{x}_k = \boldsymbol{t}_x, \tag{9}$$

where $G(w, d)$ is a distance measure satisfying the usual conditions required in the calibration paradigm (see e.g. Deville and Särndal, 1992, Section 2). Note that $\hat{\boldsymbol{t}}_{xH} = \sum_{k \in s} d_k^\circ \boldsymbol{x}_k$ is the Hartley estimator of $\boldsymbol{t}_x$ for a given $\eta$. Then $w_k^\circ = d_k^\circ F(\boldsymbol{x_k}\boldsymbol{\lambda})$, where $F(u) = g^{-1}(u)$ and $g^{-1}(\cdot)$ denotes the inverse function of $g(w, d) = \partial G(w, d)/\partial w$. The vector $\boldsymbol{\lambda}$ is determined using

$$\phi_s(\boldsymbol{\lambda}) = \sum_{k \in s} d_k^\circ [F(\boldsymbol{x}_k \boldsymbol{\lambda}) - 1] \boldsymbol{x}_k^T,$$

so that $\phi_s(\boldsymbol{\lambda}) = \boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH}$.

Given the set of constraints, different calibration estimators are obtained by using different distance measures. In many instances, numerical methods are required to solve the the minimization problem in (9). However, it is well known that, if we take the Euclidean (or $\chi^2$-statistic) type of distance function $G(w_k^\circ, d_k^\circ) = (w_k^\circ - d_k^\circ)^2/2d_k^\circ$, equivalent to the *linear* method in Deville et al.

(1993), we can obtain an analytic solution. In particular,

$$w_k^\circ = d_k^\circ(1 + \boldsymbol{x}_k\boldsymbol{\lambda}) \tag{10}$$

and, substituting this value in the calibration constraint in (9), we obtain $\boldsymbol{\lambda} = [\sum_{k\in s} d_k^\circ \boldsymbol{x}_k^T \boldsymbol{x}_k]^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})^T$. Substituting this value back into equation (10), the weights take the following form

$$w_k^\circ = d_k^\circ\Big[1 + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})\Big(\sum_{k\in s} d_k^\circ \boldsymbol{x}_k^T \boldsymbol{x}_k\Big)^{-1}\boldsymbol{x}_k^T\Big]. \tag{11}$$

In this case, estimator $\hat{Y}_{\text{CAL}}$ can be written as:

$$\begin{aligned}\hat{Y}_{\text{CAL}} &= \sum_{k\in s} w_k^\circ y_k = \hat{Y}_H + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})\hat{\boldsymbol{\beta}}^\circ \\ &= \sum_{k=1}^{N} \boldsymbol{x}_k\hat{\boldsymbol{\beta}}^\circ + \sum_{k\in s} d_k^\circ(y_k - \boldsymbol{x}_k\hat{\boldsymbol{\beta}}^\circ),\end{aligned}$$

where $\hat{\boldsymbol{\beta}}^\circ = (\sum_{k\in s} d_k^\circ \boldsymbol{x}_k^T \boldsymbol{x}_k)^{-1}(\sum_{k\in s} d_k^\circ \boldsymbol{x}_k^T y_k)$. This estimator takes the form of a *generalized regression* type estimator for dual frame surveys and will be denoted by $\hat{Y}_{\text{GREG}}$. These results are in line with those of calibration estimator in one frame surveys: Horvitz-Thompson estimators of $Y$ and $\boldsymbol{t}_x$, and in the regression coefficient are here replaced by Hartley estimators.

Now, if we take as distance function $G(\cdot)$ the Kullback-Leibler divergence defined as

$$G(w_k^\circ, d_k^\circ) = -d_k^\circ \log(w_k^\circ/d_k^\circ) + w_k^\circ - d_k^\circ, \tag{12}$$

that is Case 4 distance examined in Deville and Särndal (1992), then $F(u) = 1/(1-u)$ and numerical methods are required. It can be noted that maximizing the PEL function in (4) is equivalent to minimizing (12) given the same set of starting weights and set of constraints. This equivalence was already noted in one frame surveys by Deville (2005).

The calibration process induces a different final value for the weights which depends on both the distance measure $G(\cdot,\cdot)$ used and the benchmark constraints applied. On the other hand, given a value for $\eta$, the final set of weights does not depend on the values of the variables of interest and can be, therefore, used for all variables of interest. When a value for $\eta$ is to be computed from

the sample data, then it is essential to consider proposals based on estimators of $N_a$, $N_b$ and $N_{ab}$ as the one in, e.g., Skinner and Rao (1996) so that it is the same for all variables of interest. In the following, we consider some relevant examples of the form taken by the calibration estimator according to the auxiliary information available. Then, the theoretical properties are proven in Section 3.6.

### 3.1 $N_A$, $N_B$ and $N_{ab}$ all known

Suppose that the dimension of the three sets $N_A$, $N_B$ and $N_{ab}$ is known. Then, we can build the auxiliary vector using domain membership indicator variables, i.e.

$$\boldsymbol{x}_k = \big(\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b)\big), \quad \text{for} \quad k = 1, \ldots, N, \tag{13}$$

where $\delta_k(a) = 1$ if $k \in a$ and 0 otherwise, $\delta_k(ab) = 1$ if $k \in ab$ and 0 otherwise, $\delta_k(ba) = 1$ if $k \in ba$ and 0 otherwise and $\delta_k(b) = 1$ if $k \in b$ and 0 otherwise. In order to have final weights that can be used directly to estimate population totals as in equation (8), we will let $\boldsymbol{t}_x = (N_a, \eta N_{ab}, (1-\eta)N_{ba}, N_b)$ be the vector of known totals. In this case the calibration constraints are given by

$$\sum_{k \in s_a} w_k^\circ = N_a, \quad \sum_{k \in s_{ab}} w_k^\circ = \eta N_{ab}, \quad \sum_{k \in s_{ba}} w_k^\circ = (1-\eta) N_{ba}, \quad \sum_{k \in s_b} w_k^\circ = N_b, \tag{14}$$

and the minimization problem has an analytic solution irrespective of the distance function employed. Such solution is given by

$$w_k^\circ = \begin{cases} d_{Ak} N_a / \hat{N}_a & \text{if } k \in s_a \\ \eta\, d_{Ak} N_{ab} / \hat{N}_{ab} & \text{if } k \in s_{ab} \\ (1-\eta)\, d_{Bk} N_{ba} / \hat{N}_{ba} & \text{if } k \in s_{ba} \\ d_{Bk} N_b / \hat{N}_b & \text{if } k \in s_b \end{cases}, \tag{15}$$

where $\hat{N}_a = \sum_{k \in s_a} d_{Ak}$, $\hat{N}_{ab} = \sum_{k \in s_{ab}} d_{Ak}$, $\hat{N}_{ba} = \sum_{k \in s_{ba}} d_{Bk}$ and $\hat{N}_b = \sum_{k \in s_b} d_{Bk}$. Note that these weights provide Hájek type estimators for each domain and mirror the result provided in Deville et al. (1993) when dealing with the calibration estimator in case auxiliary information consists of known cell counts in a frequency table. Deville et al. (1993) denote this case as *complete post-stratification*, that is when all the domain sizes are known and used for calibration. Note that, given that we are estimating totals in domains using ratio type estimators, the sample size of the domains is important to avoid the introduction of possible bias in the final estimates.

### 3.2 $N_A$, $N_B$ known and $N_{ab}$ unknown

Following the terminology of Deville et al. (1993), we call the case treated in this section as *incomplete post-stratification* and we mean that not all the domain sizes are known, in particular we know only the size of frame $A$ and of frame $B$, but we don't known the size of the overlap domain $ab$. In this case, for $k = 1, \ldots, N$, we can write the vector $\boldsymbol{x}$ of auxiliary information as:

$$\boldsymbol{x}_k = (\delta_k(a) + \delta_k(ab) + \delta_k(ba), \delta_k(b) + \delta_k(ab) + \delta_k(ba)). \tag{16}$$

The vector of known totals in this case is $\boldsymbol{t}_x = (N_A, N_B)$ and we have the following calibration constraints

$$\begin{aligned} \sum_{k \in s_a} w_k^\circ + \sum_{k \in s_{ab}} w_k^\circ + \sum_{k \in s_{ba}} w_k^\circ &= N_A \\ \sum_{k \in s_b} w_k^\circ + \sum_{k \in s_{ab}} w_k^\circ + \sum_{k \in s_{ba}} w_k^\circ &= N_B, \end{aligned}$$

in which we, in some sense, calibrate on the margins. Final calibration weights are no longer independent from the distance function used and it is not possible to obtain an analytical expression unless we use the Euclidean distance. In this latter case we obtain the following estimator of $N_{ab}$:

$$\hat{N}_{ab}^w = \hat{N}_{ab,H} \frac{\hat{N}_a N_B + \hat{N}_b N_A - \hat{N}_a \hat{N}_b}{\hat{N}_a \hat{N}_B + \hat{N}_b \hat{N}_A - \hat{N}_a \hat{N}_b}, \tag{17}$$

where $\hat{N}_{ab,H} = \eta \hat{N}_{ab} + (1 - \eta)\hat{N}_{ba}$. We can note how the calibration procedure adjusts the Hartley estimator of $N_{ab}$ accounting for auxiliary information.

Rao and Wu (2010) also consider the case in which $N_{ab}$ is unknown. However, they do not estimate it from within the maximum PEL procedure, but they first estimate it by $\hat{N}_{ab,P} = \hat{\theta}\hat{N}_{ab} + (1 - \hat{\theta})\hat{N}_{ba}$, where $\hat{\theta} = v(\hat{N}_{ba})/\{v(\hat{N}_{ab}) + v(\hat{N}_{ba})\}$ and $v$ denotes variance estimates. Then, they take a *pseudo-complete post stratification* approach by suitably modifying the likelihood function.

### 3.3 Population totals for group membership indicators are known

Let the population $\mathcal{U}$ be divided into $H$ mutually exclusive groups $\mathcal{U}_h$, for $h = 1, \ldots, H$ such that $\bigcup_{h=1}^{H} \mathcal{U}_h = \mathcal{U}$ and let $\delta_k(h)$ be the indicator variable that takes value 1 if unit $k \in \mathcal{U}_h$ and 0

otherwise, for $k = 1, \ldots, N$ and $h = 1, \ldots, H$. Then, $\sum_{k=1}^{N} \delta_k(h) = N_h$ and $\sum_{h=1}^{H} N_h = N$. Now, consider the situation in which we know the population total of such indicator variables for each of the four domains, i.e. $N_{a,h} = \sum_{k \in a} \delta_k(h)$, $N_{ab,h} = \sum_{k \in ab} \delta_k(h)$, $N_{ba,h} = \sum_{k \in ba} \delta_k(h) = N_{ab,h}$, $N_{b,h} = \sum_{k \in b} \delta_k(h)$, for $h = 1, \ldots, H$. Note that $N_{a,h} = \sum_{k \in a} \delta_k(h) = \sum_{k=1}^{N} \delta_k(a)\delta_k(h)$ and similarly for the other cases. In practice, this would mean that we know, say the number of units for each of $H$ age-sex groups in the population for each of the four domains. This amount of auxiliary information of course implies that we also know the dimension of the three sets $N_A$, $N_B$ and $N_{ab}$ considered in the Section 3.1. Indeed, that is a special case of the present one.

In this case the vector of auxiliary variables is defined for $k = 1, \ldots, N$ by

$$\boldsymbol{x}_k = \{(\delta_k(a)\delta_k(h), \delta_k(ab)\delta_k(h), \delta_k(ba)\delta_k(h), \delta_k(b)\delta_k(h)\}_{h=1,\ldots,H}$$

and the vector of known totals is set to be $\boldsymbol{t}_x = \{(N_{a,h}, \eta N_{ab,h}, (1-\eta)N_{ba,h}, N_{b,h})\}_{h=1,\ldots,H}$. As in Section 3.1 the minimization problem has an analytic solution irrespective of the distance function employed. Such solution is given by

$$w_k^{\circ} = \begin{cases} d_{Ak} N_{a,h} / \hat{N}_{a,h} & \text{if } k \in \{s_a \cap \mathcal{U}_h\} \\ \eta\, d_{Ak} N_{ab,h} / \hat{N}_{ab,h} & \text{if } k \in \{s_{ab} \cap \mathcal{U}_h\} \\ (1-\eta)\, d_{Bk} N_{ba,h} / \hat{N}_{ba,h} & \text{if } k \in \{s_{ba} \cap \mathcal{U}_h\} \\ d_{Bk} N_{b,h} / \hat{N}_{b,h} & \text{if } k \in \{s_b \cap \mathcal{U}_h\} \end{cases} \quad \text{for } h = 1, \ldots, H, \tag{18}$$

where $\hat{N}_{a,h} = \sum_{k \in s_a} d_{Ak} \delta_k(h)$ and similarly for the other size estimators. This is another case of complete post-stratification. The final estimator will be more efficient than the Hartley estimator as much as groups collect units with a similar value of the variable of interest.

When, on the other side, we only know the population total in frame $A$ and in frame $B$, i.e. we do not know the distribution for the intersection domain $ab$, then we are again in a situation of incomplete post-stratification, like that of Section 3.2. Here,

$$\boldsymbol{x}_k = \{[\delta_k(a) + \delta_k(ab) + \delta_k(ba)]\delta_k(h), [\delta_k(b) + \delta_k(ab) + \delta_k(ba)]\delta_k(h)\}_{h=1\ldots,H}$$

and $\boldsymbol{t}_x = \{(N_{A,h}, N_{B,h})\}_{h=1\ldots,H}$. We have an analytic solution for the form of the weights only for the Euclidean distance case, but it does not take a simple tractable form as that considered in

Section 3.2. A similar situation arises also when, as in the case considered later in the application (Section 7), we do not know the distribution for, say, age-sex groups, but we know only the total for age and the total for sex, in each of the two frames $A$ and $B$. This is another example of incomplete post-stratification, that employs a form of raking (depending on the distance function employed) to obtain the final set of weights (see also examples in Section 4).

### 3.4 $N_A$, $N_B$, $N_{ab}$ known and $X_A$ known

Suppose that we know not only the frame sizes $N_A$, $N_B$ and $N_{ab}$, but, also the population total of an auxiliary numerical variable $x_A$ correlated to the study variable $y$ and relative to frame $A$, whose total is $X_A = \sum_{k\in\mathcal{A}} x_{Ak}$. In this case the vector of auxiliary variables is defined for $k = 1 \ldots, N$ by

$$\boldsymbol{x}_k = (\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b), [\delta_k(a) + \delta_k(ab) + \delta_k(ba)]x_{Ak})$$

and the calibration constraints are those in (14) plus

$$\sum_{k\in s_a} w_k^\circ x_{Ak} + \sum_{k\in s_{ab}} w_k^\circ x_{Ak} + \sum_{k\in s_{ba}} w_k^\circ x_{Ak} = X_A. \tag{19}$$

Again, it is not possible to obtain an analytic expression for the calibration weights unless we use the Euclidean distance for the Lagrange function. It can be shown that, in this case, the calibrated weights for $k \in s_a$ are such that

$$w_k^\circ = d_{Ak}\left[\frac{N_a}{\hat{N}_a} + \lambda(\frac{\hat{X}_a}{\hat{N}_a} - x_{Ak})\right], \tag{20}$$

where $\lambda$ is the Lagrange multiplier for the last constraint in (19) given by

$$\lambda = \frac{X_A - \hat{X}_{A,\text{Háj}}}{\hat{S}^2_{a,x} + \eta\hat{S}^2_{ab,x} + (1-\eta)\hat{S}^2_{ba,x}}$$

where $\hat{X}_{A,\text{Háj}}$ is a Hartley type estimator in which each component is estimated using the Hájek estimator, $\hat{S}^2_{a,x} = \sum_{k\in s_a} d_{ak}(x_{Ak} - \hat{X}_a/\hat{N}_a)^2$ and similarly for $\hat{S}^2_{ab,x}$ and $\hat{S}^2_{ba,x}$. Calibrated weights $w_k^\circ$ for $k \in s_{ab}$ and for $k \in s_{ba}$ are similar to those in (20) but with quantities referred to the appropriate domain, while weights for $k \in s_b$ are the same as in (15). With such weights, the

resulting calibration estimator resembles a combined regression estimator; in fact

$$\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{Háj}} + (X_A - \hat{X}_{A,\text{Háj}})\hat{\beta}_A$$

where $\hat{Y}_{\text{Háj}}$ is the Hartley estimator of $Y$ in which each component is estimated by its Hájek estimator, while

$$\hat{\beta}_A = \frac{\hat{S}_{a,xy} + \eta\hat{S}_{ab,xy} + (1-\eta)\hat{S}_{ba,xy}}{\hat{S}^2_{a,x} + \eta\hat{S}^2_{ab,x} + (1-\eta)\hat{S}^2_{ba,x}},$$

with $\hat{S}_{a,xy} = \sum_{k\in s_a} d_{ak}(x_{Ak} - \hat{X}_a/\hat{N}_a)(y_k - \hat{Y}_a/\hat{N}_a)$ and similarly for $\hat{S}_{ab,xy}$ and $\hat{S}_{ba,xy}$.

### 3.5 Other examples

The cases previously discussed are only a few examples of the very many possible ones that can be treated with calibration. The calibration approach is very flexible and can also handle both indicator and numerical variables simultaneously. Next we provide some details on how to construct the auxiliary vector and the vector of control totals for other interesting cases in practice; some of these cases will be used in the simulation study and in the application.

**$N_A$, $N_B$, $N_{ab}$ known and $X$ known.** Suppose that we know the frame sizes $N_A$, $N_B$ and $N_{ab}$, and let the population total of an auxiliary numerical variable be available for the whole population $X = \sum_{k=1}^{N} x_k$ and not only for frame $A$ as in the previous section. The auxiliary vector is thus $\boldsymbol{x}_k = (\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b), x_k)$ and the calibration constraints are those in (14) plus $\sum_{k\in s} w^\circ_k x_k = X$.

**$N_A$, $N_B$, known and $X_A$ and $Z_B$ known.** Suppose that we know the frame sizes $N_A$, $N_B$ and the population total of an auxiliary numerical variable $x_A$ relative to frame $A$, whose total is $X_A = \sum_{k\in\mathcal{A}} x_{Ak}$ and the population total of another auxiliary numerical variable $z_B$ relative to frame $B$, whose total is $Z_B = \sum_{k\in\mathcal{B}} z_B$. The auxiliary vector is

$$\boldsymbol{x}_k = (\delta_k(a)+\delta_k(ab)+\delta_k(ba), \delta_k(b)+\delta_k(ab)+\delta_k(ba), [\delta_k(a)+\delta_k(ab)+\delta_k(ba)]x_{Ak}, [\delta_k(b)+\delta_k(ab)+\delta_k(ba)]z_{Bk})$$

and the vector of known totals in this case is $\boldsymbol{t}_x = (N_A, N_B, X_A, Z_B)$, which allows us to write

the following calibration constraints

$$\begin{aligned}
&\sum_{k\in s_a} w_k^\circ + \sum_{k\in s_{ab}} w_k^\circ + \sum_{k\in s_{ba}} w_k^\circ = N_A \\
&\sum_{k\in s_b} w_k^\circ + \sum_{k\in s_{ab}} w_k^\circ + \sum_{k\in s_{ba}} w_k^\circ = N_B, \\
&\sum_{k\in s_a} w_k^\circ x_{Ak} + \sum_{k\in s_{ab}} w_k^\circ x_{Ak} + \sum_{k\in s_{ba}} w_k^\circ x_{Ak} = X_A \\
&\sum_{k\in s_b} w_k^\circ z_{Bk} + \sum_{k\in s_{ab}} w_k^\circ z_{Bk} + \sum_{k\in s_{ba}} w_k^\circ z_{Bk} = Z_B.
\end{aligned} \tag{21}$$

**$N_A$, $N_B$, $N_{ab}$ known and $X_A$, $X_B$ known.** When we know the frame sizes $N_A$, $N_B$ and $N_{ab}$ and the population totals of the same auxiliary variable $x$ in the two frames $X_A$ and $X_B$, the auxiliary vector is

$$\boldsymbol{x}_k = (\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b), [\delta_k(a) + \delta_k(ab) + \delta_k(ba)]x_k, [\delta_k(b) + \delta_k(ab) + \delta_k(ba)]x_k)$$

and the vector of known totals in this case is $\boldsymbol{t}_x = (N_a, N_{ab}, N_{ba}, N_b, X_A, X_B)$.

### 3.6 Asymptotic properties of $\hat{Y}_{\mathbf{CAL}}$

To show the asymptotic properties of the general calibration estimator we adapt and place ourselves in the asymptotic framework of Isaki and Fuller (1982), in which the dual-frame finite population $\mathcal{U}$ and the sampling designs $p_A(\cdot)$ and $p_B(\cdot)$ are embedded into a sequence of such populations and designs indexed by $N$, $\{\mathcal{U}_N, p_{A_N}(\cdot), p_{B_N}(\cdot)\}$, with $N \to \infty$. We will assume therefore, that $N_{A_N}$ and $N_{B_N}$ tend to infinity and that also $n_{A_N}$ and $n_{B_N}$ tend to infinity as $N \to \infty$. We will further assume that $N_a > 0$ and $N_b > 0$. In addition $n_{A_N}/n_N \to c_1 \in (0,1)$, where $n_N = n_{A_N} + n_{B_N}$, $N_a/N_A \to c_2 \in (0,1)$, $N_b/N_B \to c_3 \in (0,1)$ as $N \to \infty$. Subscript $N$ may be dropped for ease of notation, although all limiting processes are understood as $N \to \infty$. Stochastic orders $O_p(\cdot)$ and $o_p(\cdot)$ are with respect to the aforementioned sequences of designs. The constant $\eta \in (0,1)$ is kept fixed over repeated sampling. In order to prove our results, we make the following technical assumptions.

**A1.** *Let* $\boldsymbol{B}_U = (\sum_{k=1}^N \boldsymbol{x}_k^T \boldsymbol{x}_k)^{-1} \sum_{k=1}^N \boldsymbol{x}_k^T y_k$. *Assume that* $\boldsymbol{B} = \lim_{N\to\infty} \boldsymbol{B}_U$ *exists; the distribution of* $\boldsymbol{x}_k$ *and of* $y_k$*, and the sampling designs are such that* $\sum_{k=1}^N \boldsymbol{x}_k^T \boldsymbol{x}_k$ *is consistently estimated by* $\sum_{k\in s} d_k^\circ \boldsymbol{x}_k^T \boldsymbol{x}_k$ *and* $\sum_{k=1}^N \boldsymbol{x}_k^T y_k$ *is consistently estimated by* $\sum_{k\in s} d_k^\circ \boldsymbol{x}_k^T y_k$.

**A2.** *The limiting design covariance matrix of the normalized Hartley estimators,*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \Sigma_{yy} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}^T & \boldsymbol{\Sigma}_{xx} \end{bmatrix} = \lim_{N\to\infty} \frac{n_N}{N^2} \begin{bmatrix} V(\hat{Y}_H) & \boldsymbol{C}(\hat{\boldsymbol{t}}_{xH}, \hat{Y}_H) \\ \boldsymbol{C}(\hat{\boldsymbol{t}}_{xH}, \hat{Y}_H)^T & \boldsymbol{V}(\hat{\boldsymbol{t}}_{xH}) \end{bmatrix}$$

*is positive defined.*

**A3.** *The normalized Hartley estimators of* $\boldsymbol{t}_x$ *and* $Y$ *are such that a central limit theorem holds:*

$$\frac{\sqrt{n_N}}{N} \begin{bmatrix} \sum_{k\in s} d_k^\circ y_k - Y \\ \sum_{k\in s} d_k^\circ \boldsymbol{x}_k^T - \boldsymbol{t}_x^T \end{bmatrix} \xrightarrow{\mathcal{L}} N(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

**A4.** *The estimated covariance matrix for the Hartley estimator is design consistent in the sense that*

$$\frac{n_N}{N^2} \begin{bmatrix} v(\hat{Y}_H) & \boldsymbol{c}(\hat{\boldsymbol{t}}_{xH}, \hat{Y}_H) \\ \boldsymbol{c}(\hat{\boldsymbol{t}}_{xH}, \hat{Y}_H)^T & \boldsymbol{v}(\hat{\boldsymbol{t}}_{xH}) \end{bmatrix} - \boldsymbol{\Sigma} = o_p(1),$$

*where* $v(\hat{Y}_H) = v(\hat{Y}_a + \eta \hat{Y}_{ab}) + v((1-\eta)\hat{Y}_{ba} + \hat{Y}_b)$ *and similarly for the others.*

We will first state the properties of $\hat{Y}_{\text{CAL}}$ for the Euclidean distance, i.e. $\hat{Y}_{\text{GREG}}$, and then show the convergence for a general distance function. The following theorem shows that $\hat{Y}_{\text{GREG}}$ is design consistent, and provides its asymptotic distribution.

**Theorem 1.** *Under assumptions A1–A3,* $\hat{Y}_{GREG}$ *is design* $\sqrt{n_N}$*-consistent for* $Y$ *in the sense that,*

$$\hat{Y}_{GREG} - Y = O_p(N n_N^{-1/2})$$

*and has the following asymptotic distribution*

$$\frac{\hat{Y}_{GREG} - Y}{\sqrt{V_\infty(\hat{Y}_{GREG})}} \xrightarrow{\mathcal{L}} N(0,1)$$

*where* $V_\infty(\hat{Y}_{GREG}) = V(\hat{t}_{eH})$ *and* $\hat{t}_{eH} = \sum_{k\in s} d_k^\circ e_k$ *is the Hartley estimator of the population total of the "census"-level residuals* $e_k = y_k - \boldsymbol{x}_k \boldsymbol{B}_U$.

**Proof.** See the Appendix.

□

A design unbiased variance estimator is available for the Horvitz-Thompson estimator for many designs, and therefore for the Hartley estimator for a given $\eta$. The following theorem shows that, in these cases, it is possible to construct a design consistent estimator for the variance of the asymptotic distribution $V_\infty(\hat{Y}_{\text{GREG}})$ obtained in Theorem 1.

**Theorem 2.** *Let $\hat{e}_k = y_k - \boldsymbol{x}_k\hat{\boldsymbol{\beta}}^\circ$. Then, under assumptions A1, A2 and A4*

$$\begin{aligned} v(\hat{Y}_{GREG}) = v(\hat{t}_{\hat{e}H}) &= v\Big(\sum_{k\in s_a} d_k\hat{e}_k + \eta\sum_{k\in s_{ab}} d_k\hat{e}_k\Big) + v\Big((1-\eta)\sum_{k\in s_{ba}} d_k\hat{e}_k + \sum_{k\in s_b} d_k\hat{e}_k\Big) = \\ &= V(\hat{t}_{eH}) + o_p(N^2n_N^{-1}). \end{aligned}$$

**Proof.** See the Appendix.

□

From Theorem 2 we can derive an asymptotic distribution result using the estimated variance as stated in the following corollary.

**Corollary 1.** *Under assumptions A1–A4, $\hat{Y}_{GREG}$ is such that*

$$\frac{\hat{Y}_{GREG} - Y}{\sqrt{v(\hat{Y}_{GREG})}} \xrightarrow{\mathcal{L}} N(0,1).$$

Now we establish the asymptotic equivalence between $\hat{Y}_{\text{CAL}}$ and $\hat{Y}_{\text{GREG}}$. To this end, we further make the following assumptions (see Section 2 Deville and Särndal, 1992).

**A5.** *$\phi_s(\boldsymbol{\lambda})$ is defined on $C = \bigcap_{k=1}^N\{\boldsymbol{\lambda} : \boldsymbol{x}_k\boldsymbol{\lambda} \in Im_k(d_k^\circ)\}$. $C$ is an open neighborhood of $\mathbf{0}$.*

**A6.** *As $N \to \infty$, $\max||\boldsymbol{x}_k|| = M < \infty$, $k = 1,\ldots,N$, and $\max F_k''(0) = M' < \infty$, where $F_k''(\cdot)$ is the second derivative of $F_k(\cdot)$.*

**Theorem 3.** *Under assumptions A1–A3 and A5–A6*

$$\hat{Y}_{CAL} - \hat{Y}_{GREG} = O_p(Nn_N^{-1}).$$

**Proof.** See the Appendix.

□

**Corollary 2.** *Under A1–A6* $\hat{Y}_{CAL}$ *is such that*

$$\frac{\hat{Y}_{CAL} - Y}{\sqrt{v(\hat{Y}_{GREG})}} \xrightarrow{\mathcal{L}} N(0,1).$$

# 4 Calibration estimation: single-frame methods

In those situations in which we know the inclusion probability of the units in the sample under both sampling designs, then we can account for it in a calibration framework employing *single-frame* estimators (Kalton and Anderson, 1986; Bankier, 1986; Skinner, 1991). The calibration estimator in this single-frame approach is given by $\hat{Y}^{\mathrm{S}}_{\mathrm{CAL}} = \sum_{k\in s} w^{\star}_k y_k$ where weights $w^{\star}_k$ are such that

$$\min \sum_{k\in s} G(w^{\star}_k, d^{\star}_k) \quad \text{s.t.} \quad \sum_{k\in s} w^{\star}_k \boldsymbol{x}_k = \boldsymbol{t}_x.$$

A general solution to the minimization problem is given by $w^{\star}_k = d^{\star}_k F(\boldsymbol{x}_k \boldsymbol{\lambda})$. Note the only difference with equation (9) is the starting basic design weight. Note that calibration can handle the case in which $(1/d_{Ak} + 1/d_{Bk}) \geq 1$ for some units $k$ and, therefore, the basic weights are smaller than 1.

If we take the Euclidean distance function, the calibration weights obtained from the minimization procedure are given by $w^{\star}_k = d^{\star}_k(1 + \boldsymbol{x}_k \boldsymbol{\lambda})$ with $\boldsymbol{\lambda} = (\sum_{k\in s} d^{\star}_k \boldsymbol{x}^T_k \boldsymbol{x}_k)^{-1}(\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xS})^T$ and $\hat{\boldsymbol{t}}_{xS} = \sum_{k\in s} d^{\star}_k \boldsymbol{x}_k$, i.e. the single-frame estimator for the total $\boldsymbol{t}_x$. As expected, the resulting calibration estimator takes a generalized regression estimator form, given by

$$\begin{aligned}
\hat{Y}^{\mathrm{S}}_{\mathrm{GREG}} &= \sum_{k\in s} d^{\star}_k y_k + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xS}) \Big( \sum_{k\in s} d^{\star}_k \boldsymbol{x}^T_k \boldsymbol{x}_k \Big)^{-1} \sum_{k\in s} d^{\star}_k \boldsymbol{x}^T_k y_k \\
&= \hat{Y}_S + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xS}) \hat{\boldsymbol{\beta}}^{\star},
\end{aligned}$$

where $\hat{\boldsymbol{\beta}}^{\star} = (\sum_{k\in s} d^{\star}_k \boldsymbol{x}^T_k \boldsymbol{x}_k)^{-1} \sum_{k\in s} d^{\star}_k \boldsymbol{x}^T_k y_k$. Under assumptions in all similar to those of Section 3.6, concerning consistency of the single-frame estimator instead of the Hartley estimator, $\hat{Y}^{\mathrm{S}}_{\mathrm{CAL}}$ can be proven to be a consistent estimator, to be asymptotically equivalent to $\hat{Y}^{\mathrm{S}}_{\mathrm{GREG}}$ and, therefore, to share its asymptotic distribution. In particular, it can be easily shown that the variance of

their asymptotic distribution is given by $V_\infty(\hat{Y}^{\mathrm{S}}_{\mathrm{GREG}}) = V(\hat{t}_{eS}) = V(\sum_{k\in s} d^\star_k e_k)$ and it can be consistently estimated using $v(\hat{t}_{\hat{e}S}) = v(\sum_{k\in s} d^\star_k \hat{e}_k)$, where $\hat{e}_k = y_k - \boldsymbol{x}_k\hat{\boldsymbol{\beta}}^\star$. Of course here, as in the dual frame context previously explored, variance estimators alternative to the one considered here based on the linearization technique and proposed in the literature to estimate the variance of the calibration estimator can be considered as well, once the set of basic design weights are properly adjusted for (e.g. those based on resampling methods or on empirical likelihood methods). We will consider Jackknife later in Section 5.

Given that in the single-frame approach each unit in the overlap domain has a weight that accounts for the expected number of times it can be selected in the sample, care should be placed in the definition of the auxiliary variable vector. In particular, in the case $N_A$, $N_B$ and $N_{ab}$ are all known, $\boldsymbol{x}_k = (\delta_k(a), \delta_k(ab) + \delta_k(ba), \delta_k(b))$, and, therefore, $\boldsymbol{t}_x = (N_a, N_{ab}, N_b)$. As in Section 3.1, the final solution does not depend on the choice of the distance function and calibrated weights take the Hájek form

$$w^\star_k = \begin{cases} d_{Ak}N_a/\hat{N}_a & \text{if } k \in s_a \\ (1/d_{Ak} + 1/d_{Bk})^{-1}N_{ab}/\hat{N}_{abS} & \text{if } k \in s_{ab}\cup s_{ba} \\ d_{Bk}N_b/\hat{N}_b & \text{if } k \in s_b \end{cases} .$$

where $\hat{N}_{abS} = \sum_{k\in s_{ab}\cup s_{ba}}(1/d_{Ak} + 1/d_{Bk})^{-1}$. Similarly, if $N_A$, $N_B$, $N_{ab}$, and $X_A$ are known, then $\boldsymbol{x}_k = (\delta_k(a), \delta_k(ab) + \delta_k(ba), \delta_k(b), [\delta_k(a) + \delta_k(ab) + \delta_k(ba)]x_{Ak})$, and $\boldsymbol{t}_x = (N_a, N_{ab}, N_b, X_A)$.

When, on the other hand, only $N_A$, $N_B$ are known, an interesting equivalence arises. In this case the auxiliary vector is defined as in (16) and final weights depend on the distance function employed. If we consider the Case 2 distance proposed in Deville and Särndal (1992), i.e.

$$G(w^\star_k, d^\star_k) = w^\star_k \log(w^\star_k/d^\star_k) - w^\star_k + d^\star_k,$$

and the particular case of simple random sampling in both frames, we obtain that $\hat{N}^w_{ab} = \sum_{k\in s_{ab}\cup s_{ba}} w^\star_k = \hat{N}^{RR}_{ab}$, where $\hat{N}^{RR}_{ab}$ is the overlap dimension estimator obtained by Skinner (1991) using Raking Ratio as the smallest root of the quadratic equation

$$\hat{N}_{abS}\, t^2 - [\hat{N}_{abS}(N_A + N_B) + (\hat{N}_a\hat{N}_{ab})n_a n_b]t + \hat{N}_{abS}N_A N_B = 0.$$

In this case the single frame calibration estimator provides a simple tool to extend such Raking Ratio estimator to general sampling designs by simply plugging in different basic design weights $d_k^\star$, and to more composite auxiliary information settings.

# 5 Jackknife estimation of variance

In this section we explore the possibility of using Jackknife to estimate the variance of the proposed calibration estimators (see e.g. Wolter, 2003, for an introduction to Jackknife methods). Dual-frame or single-frame calibration estimators will be denoted by $\hat{Y}_c$ for short in this section.

If we consider a non stratified design, the Jackknife estimator for the variance of $\hat{Y}_c$ may be given by

$$v_J(\hat{Y}_c) = \frac{n_A - 1}{n_A} \sum_{i \in s_A} (\hat{Y}_c^A(i) - \overline{Y}_c^A)^2 + \frac{n_B - 1}{n_B} \sum_{j \in s_B} (\hat{Y}_c^B(j) - \overline{Y}_c^B)^2 \tag{22}$$

where $\hat{Y}_c^A(i)$ is the value taken by estimator $\hat{Y}_c$ after dropping unit $i$ from $s_A$ and $\overline{Y}_c^A$ is the average of $\hat{Y}_c^A(i)$ values; $\hat{Y}_c^B(j)$ and $\overline{Y}_c^B$ are defined similarly. This Jackknife estimator of the variance is conservative (upward biased) in finite populations when sampling without replacement (see Wolter, 2003, Section 4.3.4). To overcome this issue, an approximate finite-population correction is employed. Then, the new Jackknife estimator of variance $v_J^*(\hat{Y}_c)$ is obtained by replacing $\hat{Y}_c^A(i)$ in (22) with $\hat{Y}_c^{A*}(i) = \hat{Y}_c + \sqrt{1 - \overline{\pi}_A}(\hat{Y}_c^A(i) - \hat{Y}_c)$, where $\overline{\pi}_A = \sum_{i \in s_A} \pi_{iA}/n_A$.

In the case of a stratified design in both frames, let frame $A$ be divided into $H$ strata and let stratum $h$ has $N_{Ah}$ observation units of which $n_{Ah}$ are sampled. Similarly, frame $B$ has $L$ strata, the stratum $l$ has $N_{Bl}$ observation units of which $n_{Bl}$ are sampled. Then, a Jackknife variance estimator of $\hat{Y}_c$ is given by

$$v_J^{st}(\hat{Y}_c) = \sum_{h=1}^{H} \frac{n_{Ah} - 1}{n_{Ah}} \sum_{i \in s_{Ah}} (\hat{Y}_c^A(hi) - \overline{Y}_c^{Ah})^2 + \sum_{l=1}^{L} \frac{n_{Bl} - 1}{n_{Bl}} \sum_{j \in s_{Bl}} (\hat{Y}_c^B(lj) - \overline{Y}_c^{Bl})^2, \tag{23}$$

where $\hat{Y}_c^A(hi)$ is the value taken by estimator $\hat{Y}_c$ after dropping unit $i$ of stratum $h$ from sample $s_{Ah}$, $\overline{Y}_c^{Ah}$ is the average of these $n_{Ah}$ values; $\hat{Y}_c^B(lj)$ and $\overline{Y}_c^{Bl}$ are defined similarly. Again, we also can obtain a modified Jackknife variance estimator $v_J^{st*}(\hat{Y}_c)$ in stratified sampling using an approximate

finite-population correction in each stratum. Asymptotic results for the Jackknife estimators can be obtained using the approach presented in Lohr and Rao (2000).

# 6 Simulation studies

We conduct an extensive simulation study to analyze the performance of the proposed estimators for surveys from two-frame finite populations. Our simulations are programmed in `R` using the `sampling` package developed by Tillé and Matei (2006) to draw the samples and to build all the calibration estimators, using the algorithms developed by Wu (2005) for the PEL approach and also developing some new `R`-code to compute calibration estimators with the Kullback-Leibler distance.

The simulated population has dimension $N = 2350$. The values of the variable of interest $y$ are generated from a normal distribution $y_k \sim N(5000, 500)$, for $k = 1, \ldots, 2350$. Units are randomly assigned to the two frames, $A$ and $B$, according to three different scenarios depending on the overlap domain size $N_{ab}$. The first scenario has a *small* overlap domain size and units are assigned to domain $a$, $b$ or $ab$ depending on the values taken by a binomial random variable $g_k \sim Bi(2, 0.3)$. In particular, if $g_k = 0$ then $k \in a$, if $g_k = 1$ then $k \in b$ and if $g_k = 2$ then $k \in ab$. The resulting sizes of the two frames are $N_A$=1309 and $N_B$=1251 and, consequently, the overlap domain size is $N_{ab}$=210. The second and the third scenarios have respectively *large* and *medium* overlap domain size, depending on the values taken by $g_k \sim Bi(2, 0.5)$, but assigning units to each domain in different ways for the two scenarios. In particular, we have 0 for domain $a$, 1 for domain $ab$ and 2 for domain $b$ in the second scenario and 0 for domain $b$, 1 for domain $a$ and 2 for domain $ab$ in the third scenario. The resulting frame sizes in the second scenario are given by $N_A$=1746 and $N_B$=1790 and the overlap domain size is $N_{ab}$=1186, while for the third scenario we have $N_A$=1790, $N_B$=1164 and $N_{ab}$=604.

Units from frame $A$ are then divided for each scenario into six strata as follows:

- sc.1 - large overlap, $N_{Ah} = (535, 279, 78, 148, 101, 168)$,
- sc.2 - small overlap, $N_{Ah} = (734, 377, 116, 187, 115, 217)$,

- sc.3 - medium overlap, $N_{Ah} = (781, 375, 114, 186, 111, 223)$.

Two auxiliary variables are then generated from the values of $y$ for frame $A$ and frame $B$, respectively, that are $x_{Ak} = (y_k - e_k)/0.5$ where $e_k \sim N(500, 300)$ and $x_{B_k} = (y_k - 1 - e_k)/1.2$, where $e_k \sim N(700, 500)$, for $k = 1, \ldots, N$. The correlation coefficient with the variable of interest is given by $\rho_A = 0.859$ and $\rho_B = 0.709$, respectively.

Samples from frame $A$ are selected using stratified simple random sampling. Samples from frame $B$ are selected by means of Midzuno sampling, with inclusion probabilities proportional to variable $z_k = y_k - N(300, 200)$, for $k = 1, \ldots, N$ and having correlation $\rho = 0.929$ with the variable of interest. For each scenario, we draw four different combinations of sample sizes for frame $A$ and frame $B$, which correspond to the following number of units per stratum:

- $n_{A_{\text{small}}} = (15, 20, 15, 20, 15, 20) = 105$ and $n_{B_{\text{small}}} = 135$,
- $n_{A_{\text{large}}} = (30, 40, 30, 40, 30, 40) = 210$ and $n_{B_{\text{small}}} = 135$,
- $n_{A_{\text{small}}} = (15, 20, 15, 20, 15, 20) = 105$ and $n_{B_{\text{large}}} = 270$,
- $n_{A_{\text{large}}} = (30, 40, 30, 40, 30, 40) = 210$ and $n_{B_{\text{large}}} = 270$.

This makes a $3 \times 2 \times 2$ design for the simulation study. For each of the 12 settings, we compute four point calibration estimators of the population total $Y$ using both the single-frame and the dual-frame approach and using four different kinds of distance functions: Euclidean, Raking, Logit and Kullback-Leibler (corresponding to the three methods considered in Deville et al., 1993, and implemented in the `sampling package`, and the distance measure close to the PEL approach, respectively). For each estimator we examine four different types of auxiliary information:

(1) $N_A$, $N_B$, $N_{ab}$ all known,

(2) $N_A$, $N_B$ known and $N_{ab}$ unknown,

(3) $N_A$, $N_B$, $N_{ab}$, $X_A$, $X_B$ all known,

(4) $N_A$, $N_B$, $X_A$, $X_B$ all known and $N_{ab}$ unknown.

We compute also the Hartley estimator (HAR), the Pseudo Maximum Likelihood estimator (PML, Skinner and Rao, 1996) when $N_{ab}$ is unknown, the single frame estimator (SF, Bankier, 1986; Kalton and Anderson, 1986) and the Raking Ratio estimator (SFRR, Skinner, 1991) for the purpose of comparison. When needed the value of $\eta$ has been estimated using

$$\hat{\eta} = N_a N_B v(\hat{N}_{ba}) / \left[ N_b N_A v(\hat{N}_{ab}) + N_a N_B v(\hat{N}_{ba}) \right], \tag{24}$$

(see Lohr and Rao, 2000). Note that this choice allows for the computation of a single set of weights for all variables of interest. For each estimator, we compute the percent relative bias RB% = $E_{MC}(\hat{Y} - Y)/Y * 100$, the percent relative mean squared error RMSE% = $E_{MC}[(\hat{Y} - Y)^2]/Y^2 * 100$ and the percent gain in efficiency GE% = $(1 - \text{RMSE}/\text{RMSE}_{SF}) * 100$ over the single frame estimator $SF$, based on 1000 simulation runs.

Tables 1 to 3 report results, one for each scenario, relative to the case in which $n_{A_{\text{small}}}$ and $n_{B_{\text{small}}}$, i.e. they are relatively smaller. The other cases are not reported since changing the sample size does not change the trend of the results. From these tables we can see that relative biases are negligible in all cases, as expected from theoretical results. In terms of RMSE%, other things being equal, single-frame estimators are more efficient than dual-frame estimators, and this can be explained by the extra-information they incorporate in the estimation process. Given a particular type of auxiliary information, it makes a little difference in terms of efficiency which distance metric we use in the calibration approach, and this is again in line with literature on the topic.

The performance in terms of efficiency of the estimators is essentially driven by the set of auxiliary variables employed, where type (3) – $N_A$, $N_B$, $N_{ab}$, $X_A$, $X_B$ all known – is the most effective as expected. In fact, the strong correlation between the study variable $y$ and the auxiliary variables $x_A$ and $x_B$ contributes in making estimates more accurate. It is interesting to note, however, that the performance of calibration estimators in setting (4) – $N_A$, $N_B$, $X_A$, $X_B$ all known and $N_{ab}$ unknown – is closer to that of setting (2) – $N_A$, $N_B$ known and $N_{ab}$ unknown – than that of setting (3), by this providing evidence of the importance of knowing the dimension of the overlap domain $N_{ab}$. This behavior becomes more clear as the overlap domain size becomes

larger (Scenarios 3 and 2).

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

As discussed above, if we consider the calibration estimator with the Kullback-Leiber distance (CAL-KL) and we add the constraint induced by the common domain mean as in equation (5), we obtain the PEL estimator proposed by Rao and Wu (2010). To evaluate the effect of including such restriction in the calibration process, we have also computed the CAL-EUC and the CAL-KL estimators that include this new restriction (overlap restriction) in all scenarios and under two particular types of auxiliary information: (1) and (3). Table 4 reports the results from this experiment. It can be noted that in this simulation study, the inclusion of this extra constraint provides little or no improvement over classical calibration. In particular, in case (1) calibration estimators without restriction work a little better than the estimators that include the restriction, while in case (3) this behavior is reversed. Note also that using this extra constraint comes at the price of having a final estimator non-linear in $y$ and, therefore, would require different sets of weights for different variables of interest. Therefore, when used in large scale surveys one may want to choose a subset of variables of interest to enter such extra benchmark constraints and then use the final set of weights for all computations.

[Table 4 about here.]

We now turn to the construction of confidence intervals for $Y$. We obtain the 95% confidence intervals based on a normal distribution and the two proposed variance estimators: linearization based $v(\hat{Y}_{GREG})$ from Theorem 2 and Corollary 2 and Jackknife as in equation (23) with finite-population correction. Table 5 shows the average length of 95% confidence intervals, the empirical

coverage probability, the inferior and the superior tail error rates. For space reason, only some cases and some sample sizes are included.

From Table 5 we can observe that coverage probability is high (greater than 93%) for all sample sizes and all scenarios. We also observe that the intervals based on the linearization variance tend to provide empirical coverages larger than the nominal ones, while Jackknife based intervals provide coverage closer to the nominal. Jackknife intervals are also shorter and the length difference is significant in some cases (e.g. large overlap size together with the $X_A$ and $X_B$ information). The worse performance of the linearization based intervals may be due to the fact that sample sizes in some strata are too small.

[Table 5 about here.]

# 7 Application

IESA, the Institute for Advanced Social Studies of Spain conducted a survey between January, 14th and February, 13th 2011 on the perception of culture in the Spanish region of Andalusia (Barometer of Culture of Andalusia - BACU). It is based on a sample drawn from two frames: landline phone frame ($A$, $N_A$ = 5,064,304) and a mobile phone frame ($B$, $N_B$ = 5,875,280). The overlap domain size is known to have dimension $N_{ab}$ = 4,421,042.

From frame $A$ a stratified random sample without replacement of dimension $n_A = 641$ was selected, where strata are made by eight geographical regions. Strata population sizes in frame $A$ are $N_{Ah}$ = (274128, 919124, 463008, 502450, 237183, 441936, 856392, 1370083) and the corresponding strata sample sizes are $n_{Ah}$ = (53, 99, 66, 62, 38, 49, 131, 143). From frame $B$ a simple random sample without replacement of size $n_B = 177$ was drawn. Sample sizes for each frame were determined so as to minimize the cost of the survey.

Among the several topics of interest in the survey, there is also the interest to estimate the percentage of undecided citizens on next political elections. As auxiliary variables there are available sex and age (in two categories, under 45 or over); both variables are observed in both frames and

their totals are known for each of the two frames $A$ and $B$.

We compare estimates of the mean of such binary variable of interest without using any auxiliary information, using the auxiliary information provided by the sizes of the frames and overlap domains, and also using additional auxiliary information from age and sex. Results are reported in Table 6. In particular, without auxiliary information, under the dual frame approach, we compute the Hartley estimator (HAR) estimating $\eta$ as in (24) and, under the single frame approach, we compute Kalton-Anderson's (SF) estimator. Dual frame pseudo-maximum likelihood (PML) estimator and the single frame raking ratio estimation (SFRR) are also computed. Calibration estimators using two levels of auxiliary information and four distances, are also reported in Table 6. The confidence intervals (and their length) based on Jackknife variance estimation are included as well.

From Table 6 we observe that the inclusion of auxiliary information provides estimates with shorter confidence intervals. This is particularly true for when using calibration on population domains, sex and age under the single frame approach. Calibration estimates (including SFRR) are all similar, and this is particularly true when comparing values within the single and the dual frame framework. However, including all available auxiliary information, both in terms of the design – hence using the single frame approach – and of population counts we obtain the best empirical performance and an estimate that is, nonetheless, coherent with the others.

[Table 6 about here.]

# 8 Conclusions

In the last years multiple-frame surveys have significantly attracted attention in survey methodology and applications. The use of more then one frame helps statisticians to obtain more reliable estimates for finite population totals or means. Incorporating available auxiliary population information at different levels also contributes to obtain more accurate estimates. In this work we have discussed the extension of the calibration framework to estimation from dual frame surveys.

Definition of the auxiliary variables and benchmark constraints have been discussed under both the single and the dual frame approach. Some of the estimators already proposed in the literature have been shown to belong to this class of calibration estimators. The cases discussed in Section 3 are only a few examples of the very many possible ones that can be treated with calibration. The calibration approach is very flexible and wide spread for one frame surveys. We wanted to import such flexibility in the field of two frame surveys.

Estimators belonging to this class have been proven to be design consistent under mild assumptions and their asymptotic distribution has been obtained. Variance estimation has been proposed under the linearization and the Jackknife framework. Results from the extensive simulation study support theoretical findings and show that, given a set of auxiliary variables, the choice of a distance function makes little difference in terms of efficiency, as it is the case also in one frame surveys. In addition, it is well known that calibration based on the Euclidean distance function can produce negative weights whilst calibration based on the Kullback-Leibler divergence or on other distance functions considered in this paper ensures always positive weights. In this paper, we have found that in the application, the calibration estimator with Euclidean distance does not give negative weights. In the simulation study, on the other hand, among the 24,000 samples (4 cases $\times$ 3 scenarios $\times$ 2 sample sizes $\times$ 1000 replicates), the calibration estimator with Euclidean distance gives negative weights in only 6 cases in Single Frame (under scenario 1, with a relatively smaller sample size and with auxiliary information of type (3) and (4)), while DF in 981 cases (in all scenarios, when with auxiliary information of type (4)).

Calibration estimation from dual frame surveys can be implemented easily using existing software for one frame populations, as for the application on data from the BACU survey. Note that calibration weights can be applied to all variables of interest. In fact, they do not depend on the value taken by the variable of interest. This is particularly valuable, because in this way calibration estimators give internal consistency. With repeated surveys, the simplicity and transparency of a fixed-weight estimator may be preferred. Fixed-weight adjustments may make year-to-year

comparisons easier in an annual survey, where the domain proportions are relatively constant over time. Standard survey software may then be used to estimate population totals using the modified weights.

The proposed calibration estimators assume that the control totals are values known without sampling errors. However, these control totals can themselves be estimated from other surveys. Calibration can be applied similarly with those estimated controls but in this case the variance estimator need to take into account such extra variation when we use estimates of totals. To obtain the variance estimator when the controls are estimated a possibility is to use the result of Section 9 in Berger et al. (2009) for each sample $s_A$ and $s_B$ separately. These Authors obtain a variance estimator of the calibration estimator that takes into account the randomness of multiple estimates controls.

The extension to more than two frames is under study as well. One important issue when dealing with more than two frames is that of using a proper notation (see Lohr and Rao, 2006; Singh and Mecatti, 2011). A first simple way around is the one, also considered in Rao and Wu (2010), in which weights from the multiplicity estimator of Mecatti (2007) are used as starting weights and calibration is applied straightforwardly. More complicated is the issue of accounting for different levels of frame information, although we believe that Singh and Mecatti (2011) may provide a good starting point. In addition, note that, given that with calibration estimation we are often estimating totals in domains using ratio type estimators (like with post-stratification), the sample size of the domains is important to avoid the introduction of possible bias in the final estimates. This issue becomes particularly relevant when moving to more than two frames. In this case, domains may easily become small areas and model based techniques could be enforced to fully exploit auxiliary information.

## Acknowledgements

The Authors are grateful to Manuel Trujillo (IESA) for providing data and information about the Barometer of Culture of Andalusia survey and to Jean-Claude Deville for useful suggestions on distance metics in calibration. This Research is partially supported by Ministerio de Educación y Ciencia (grant MTM2012-35650, Spain) and by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía). The work of Ranalli has been developed partially under the support of the project PRIN-SURWEY (grant 2012F42NS8, Italy).

# A Proofs

**Proof of Theorem 1**

By assumptions A1 and A2 we have

$$\begin{aligned}\hat{Y}_{\text{GREG}} - Y &= \hat{Y}_H + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})\boldsymbol{B}_U - Y + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})(\hat{\boldsymbol{\beta}}^{\circ} - \boldsymbol{B}_U) \\ &= \hat{Y}_H + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})\boldsymbol{B}_U - Y + O_p(Nn_N^{-1/2})o_p(1).\end{aligned}$$

Now, $\hat{Y}_H + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})\boldsymbol{B}_U$ is such that a central limit theorem holds for A2 and A3, i.e.

$$\frac{\sqrt{n_N}}{N}(\hat{Y}_H + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})\boldsymbol{B}_U - Y) \xrightarrow{\mathcal{L}} N(0, \nu^2)$$

where $\nu^2 = \Sigma_{yy} - 2\boldsymbol{\Sigma}_{xy}\boldsymbol{B} + \boldsymbol{B}^T\boldsymbol{\Sigma}_{xx}\boldsymbol{B}$. Now, $N^2 n_N V(\hat{t}_{eH}) \to \nu^2$ as $N \to \infty$, so that $\hat{Y}_H + (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})\boldsymbol{B}_U - Y = O_p(Nn_N^{-1/2})$ and the result follows.

□

**Proof of Theorem 2**

Let $\tilde{y}_k = \boldsymbol{x}_k \boldsymbol{B}_U$ and $\hat{y}_k = \boldsymbol{x}_k \hat{\boldsymbol{\beta}}^\circ$. Then

$$\begin{aligned} v(\hat{t}_{\hat{e}H}) &= v(\hat{t}_{\hat{e}H} + \hat{t}_{eH} - \hat{t}_{eH}) = \\ &= v(\sum_{k\in s} d_k^\circ \hat{e}_k + \sum_{k\in s} d_k^\circ e_k - \sum_{k\in s} d_k^\circ e_k) = \\ &= v\Big(\sum_{k\in s} d_k^\circ e_k + \sum_{k\in s} d_k^\circ (y_k - \hat{y}_k - y_k + \tilde{y}_k)\Big) = \\ &= v(\hat{t}_{eH}) + v(\hat{t}_{\tilde{y}-\hat{y},H}) + 2c(\hat{t}_{eH}, \hat{t}_{\tilde{y}-\hat{y},H}). \end{aligned} \tag{A.1}$$

Now, for A1, A2 and A4, we have

1. $v(\hat{t}_{eH}) = V(\hat{t}_{eH}) + o_p(N^2 n_N^{-1})$,
2. $v(\hat{t}_{\tilde{y}-\hat{y},H}) = v(\sum_{k\in s} d_k^\circ \boldsymbol{x}_k(\boldsymbol{B}_U - \hat{\boldsymbol{\beta}}^\circ)) = (\boldsymbol{B}_U - \hat{\boldsymbol{\beta}}^\circ)^T v(\hat{\boldsymbol{t}}_{xH})(\boldsymbol{B}_U - \hat{\boldsymbol{\beta}}^\circ) = o_p(1) O_p(N^2 n_N^{-1}) o_p(1)$,
3. $c(\hat{t}_{eH}, \hat{t}_{\tilde{y}-\hat{y},H}) = c\big(\sum_{k\in s} d_k^\circ e_k, \sum_{k\in s} d_k^\circ \boldsymbol{x}_k(\boldsymbol{B}_U - \hat{\boldsymbol{\beta}}^\circ)\big) = \boldsymbol{c}(\hat{t}_{eH}, \hat{\boldsymbol{t}}_{xH})(\boldsymbol{B}_U - \hat{\boldsymbol{\beta}}^\circ) = O_p(N^2 n_N^{-1}) o_p(1)$.

□

**Proof of Theorem 3**

Using Result 3 in Deville and Särndal (1992)

$$\boldsymbol{\lambda} = \Big(\sum_{k\in s} d_k^\circ \, \boldsymbol{x}_k^T \boldsymbol{x}_k\Big)^{-1} (\boldsymbol{t}_x - \hat{\boldsymbol{t}}_{xH})^T + O_p(n_N^{-1}),$$

$w_k = d_k^\circ F(\boldsymbol{x}_k \boldsymbol{\lambda}) =: d_k^\circ (1 + \boldsymbol{x}_k \boldsymbol{\lambda}) + \epsilon_k(\boldsymbol{x}_k \boldsymbol{\lambda})$. Assumption A6 ensures that $\epsilon_k(u) = O_p(u^2)$, therefore

$$\hat{Y}_{\text{CAL}} = \hat{Y}_{\text{GREG}} + O_p(N n_N^{-1}) + O_p(N n_N^{-2}).$$

□

# 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.

Berger, Y. G., Muñoz, J. F., and Rancourt, E. (2009). Variance estimation of survey estimates calibrated on estimated control totals. An application to the extended regression estimator and the regression composite estimator. *Computational Statistics & Data Analysis*, 53(7):2596–2604.

Deville, J. C. (2005). Calibration: past, present and future? *Paper presented at the Workshop on "Calibration tools for survey statisticians"*, Neuchâtel, 8-9 September.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Deville, J. C., Särndal, C. E., and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88:1013–1020.

Fuller, W. A. and Burmeister, L. F. (1972). Estimators for samples selected from two overlapping frames. *Proceedings of social science section of The American Statistical Asociation.*

Hartley, H. O. (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, pages 203–206.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Kalton, G. and Anderson, D. W. (1986). Sampling rare populations. *Journal of the Royal Statistical Society. Series A (General)*, 149:65–82.

Lohr, S. L. (2009). Multiple-frame surveys. *Handbook of Statistics*, 29:71–88.

Lohr, S. L. and Rao, J. N. K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95:271–280.

Lohr, S. L. and Rao, J. N. K. (2006). Estimation in multiple-frame surveys. *Journal of the American Statistical Association*, 101(475):1019–1030.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey methodology*, 33(2):151–157.

Rao, J. N. K. and Wu, C. (2010). Pseudo–empirical likelihood inference for multiple frame surveys. *Journal of the American Statistical Association*, 105(492):1494–1503.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2):99–119.

Singh, A. C. and Mecatti, F. (2011). Generalized multiplicity-adjusted horvitz-thompson estimation as a unified approach to multiple frame surveys. *Journal of official statistics*, 27(4):1 – 19.

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. and Rao, J. N. K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91:349–356.

Tillé, Y. and Matei, A. (2006). The R package sampling, a software tool for training in official statistics and survey sampling. *Proceedings in Computational Statistics, COMPSTAT'06, Physica-Verlag/Springer*, pages 1473–1482.

Wolter, K. (2003). *Introduction to Variance estimation*. Springer-Verlag, New York.

Wu, C. (2005). Algorithms and r codes for the pseudo empirical likelihood method in survey sampling. *Survey Methodology*, 31(2):239.

Wu, C. and Rao, J. N. K. (2006). Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, 34(3):359–375.

Table 1: Scenario 1: *small* overlap domain size – $N_{ab}$=210, $N_A$=1309, $N_B$=1251 – sample sizes: $n_A$= 105, $n_B$= 135.

| | *Single Frame* | | | *Dual Frame* | | |
|---|---|---|---|---|---|---|
| | RB % | 100*RMSE % | GE % | RB % | 100*RMSE % | GE % |
| | | | (1) $N_A, N_{ab}, N_B$ *known* | | | |
| CAL (*) | -0.025 | 0.511 | 87.416 | -0.021 | 0.514 | 87.336 |
| | | | (2) $N_A, N_B$ *known,* $N_{ab}$ *unknown* | | | |
| HAR | - | - | - | -0.365 | 3.658 | 9.939 |
| PML | - | - | - | 0.113 | 2.621 | 35.470 |
| SF | -0.147 | 4.062 | 0.000 | - | - | - |
| SFRR | -0.128 | 2.315 | 43.002 | - | - | - |
| CAL-EUC | -0.133 | 2.322 | 42.821 | 0.032 | 2.587 | 36.313 |
| CAL-RAK | -0.128 | 2.315 | 43.002 | 0.036 | 2.584 | 36.379 |
| CAL-LOG | -0.128 | 2.316 | 42.982 | 0.035 | 2.584 | 36.372 |
| CAL-KL | -0.122 | 2.308 | 43.180 | 0.039 | 2.581 | 36.443 |
| | | | (3) $N_A, N_{ab}, N_B, X_A, X_B$ *known* | | | |
| CAL-EUC | -0.015 | 0.196 | 95.178 | -0.013 | 0.224 | 94.497 |
| CAL-RAK | -0.015 | 0.195 | 95.195 | -0.013 | 0.222 | 94.528 |
| CAL-LOG | -0.015 | 0.195 | 95.193 | -0.013 | 0.222 | 94.526 |
| CAL-KL | -0.013 | 0.195 | 95.199 | -0.010 | 0.224 | 94.481 |
| | | | (4) $N_A, N_B, X_A, X_B$ *known,* $N_{ab}$ *unknown* | | | |
| CAL-EUC | 0.087 | 2.187 | 46.154 | 0.087 | 2.231 | 45.083 |
| CAL-RAK | 0.086 | 2.186 | 46.181 | 0.087 | 2.227 | 45.171 |
| CAL-LOG | 0.086 | 2.186 | 46.177 | 0.087 | 2.227 | 45.162 |
| CAL-KL | 0.087 | 2.185 | 46.192 | 0.088 | 2.229 | 45.132 |

(*) irrespective of the choice of the distance measure

Table 2: Scenario 2: *large* overlap domain size - $N_{ab}$=1186, $N_A$=1746, $N_B$=1790. Samples sizes: $n_A$= 105, $n_B$= 135

| | *Single Frame* | | | *Dual Frame* | | |
|---|---|---|---|---|---|---|
| | RB % | 100*RMSE % | GE % | RB % | 100*RMSE % | GE % |
| | (1) $N_A, N_{ab}, N_B$ *known* | | | | | |
| CAL (*) | 0.021 | 0.578 | 95.282 | 0.026 | 0.605 | 95.059 |
| | (2) $N_A, N_B$ *known,* $N_{ab}$ *unknown* | | | | | |
| HAR | - | - | - | -0.144 | 9.399 | 23.252 |
| PML | - | - | - | -0.029 | 7.505 | 38.717 |
| SF | -0.062 | 12.246 | 0.000 | - | - | - |
| SFRR | -0.219 | 7.251 | 40.788 | - | - | - |
| CAL-EUC | -0.327 | 7.448 | 39.178 | -0.336 | 7.765 | 36.591 |
| CAL-RAK | -0.219 | 7.251 | 40.788 | -0.240 | 7.572 | 38.170 |
| CAL-LOG | -0.231 | 7.271 | 40.624 | -0.250 | 7.591 | 38.009 |
| CAL-KL | -0.108 | 7.114 | 41.905 | -0.143 | 7.426 | 39.359 |
| | (3) $N_A, N_{ab}, N_B, X_A, X_B$ *known* | | | | | |
| CAL-EUC | 0.026 | 0.215 | 98.241 | 0.034 | 0.301 | 97.542 |
| CAL-RAK | 0.027 | 0.215 | 98.244 | 0.035 | 0.298 | 97.563 |
| CAL-LOG | 0.027 | 0.215 | 98.244 | 0.035 | 0.299 | 97.562 |
| CAL-KL | 0.030 | 0.239 | 98.047 | 0.038 | 0.312 | 97.451 |
| | (4) $N_A, N_B, X_A, X_B$ *known,* $N_{ab}$ *unknown* | | | | | |
| CAL-EUC | -0.278 | 7.259 | 40.721 | -0.271 | 7.353 | 39.956 |
| CAL-RAK | -0.278 | 7.261 | 40.706 | -0.270 | 7.353 | 39.958 |
| CAL-LOG | -0.278 | 7.261 | 40.707 | -0.270 | 7.353 | 39.956 |
| CAL-KL | -0.278 | 7.263 | 40.687 | -0.274 | 7.360 | 39.895 |

(*) irrespective of the choice of the distance measure

Table 3: Scenario 3: *medium* overlap domain size - $N_{ab}$=604, $N_A$=1790, $N_B$=1164 Samples sizes: $n_A$= 105, $n_B$= 135

| | *Single Frame* | | | *Dual Frame* | | |
|---|---|---|---|---|---|---|
| | RB % | 100*RMSE % | GE % | RB % | 100*RMSE % | GE % |
| | (1) $N_A, N_{ab}, N_B$ *known* | | | | | |
| CAL (*) | 0.006 | 0.761 | 95.036 | 0.007 | 0.779 | 94.920 |
| | (2) $N_A, N_B$ *known,* $N_{ab}$ *unknown* | | | | | |
| HAR | - | - | - | -0.016 | 13.453 | 12.268 |
| PML | - | - | - | 0.265 | 4.513 | 70.567 |
| SF | 0.213 | 15.334 | 0.000 | - | - | - |
| SFRR | 0.055 | 4.271 | 72.148 | - | - | - |
| CAL-EUC | 0.013 | 4.333 | 71.744 | 0.074 | 4.510 | 70.587 |
| CAL-RAK | 0.055 | 4.271 | 72.148 | 0.109 | 4.469 | 70.855 |
| CAL-LOG | 0.050 | 4.277 | 72.108 | 0.105 | 4.473 | 70.828 |
| CAL-KL | 0.096 | 4.230 | 72.417 | 0.144 | 4.442 | 71.032 |
| | (3) $N_A, N_{ab}, N_B, X_A, X_B$ *known* | | | | | |
| CAL-EUC | 0.004 | 0.226 | 98.527 | -0.006 | 0.356 | 97.676 |
| CAL-RAK | 0.002 | 0.226 | 98.526 | -0.007 | 0.354 | 97.694 |
| CAL-LOG | 0.002 | 0.226 | 98.526 | -0.007 | 0.354 | 97.693 |
| CAL-KL | 0.000 | 0.227 | 98.523 | -0.008 | 0.355 | 97.683 |
| | (4) $N_A, N_B, X_A, X_B$ *known,* $N_{ab}$ *unknown* | | | | | |
| CAL-EUC | 0.174 | 3.762 | 75.468 | 0.163 | 3.955 | 74.207 |
| CAL-RAK | 0.172 | 3.762 | 75.465 | 0.163 | 3.956 | 74.200 |
| CAL-LOG | 0.172 | 3.762 | 75.465 | 0.163 | 3.956 | 74.201 |
| CAL-KL | 0.170 | 3.762 | 75.468 | 0.163 | 3.974 | 74.082 |

(*) irrespective of the choice of the distance measure

Table 4: Efficiency of Kullback-Leiber (KL) and Euclidean (EUC) distance based calibration estimators *With* and *Without* overlap restriction (5) in the dual frame approach.

| | | | *With* | | | *Without* | | |
|---|---|---|---|---|---|---|---|---|
| *Scenario* | $(n_A, n_B)$ | | RB % | 100*RMSE % | GE % | RB % | 100*RMSE % | GE % |
| | | | (1) $N_A, N_B$, $N_{ab}$, *known* | | | | | |
| *Small* | (105,135) | KL | -0.019 | 0.529 | 86.972 | -0.021 | 0.514 | 87.336 |
| | | EUC | -0.020 | 0.521 | 87.171 | -0.021 | 0.514 | 87.336 |
| | (210,270) | KL | 0.017 | 0.253 | 87.260 | 0.020 | 0.244 | 87.709 |
| | | EUC | 0.020 | 0.249 | 87.446 | 0.020 | 0.244 | 87.709 |
| *Large* | (105,135) | KL | 0.021 | 0.644 | 94.742 | 0.026 | 0.605 | 95.059 |
| | | EUC | 0.025 | 0.643 | 94.750 | 0.026 | 0.605 | 95.059 |
| | (210,270) | KL | -0.004 | 0.273 | 95.224 | -0.006 | 0.258 | 95.487 |
| | | EUC | -0.003 | 0.274 | 95.211 | -0.006 | 0.258 | 95.487 |
| *Medium* | (105,135) | KL | 0.022 | 0.817 | 94.672 | 0.007 | 0.779 | 94.920 |
| | | EUC | 0.017 | 0.813 | 94.698 | 0.007 | 0.779 | 94.920 |
| | (210,270) | KL | -0.005 | 0.385 | 94.341 | -0.002 | 0.367 | 94.606 |
| | | EUC | -0.004 | 0.385 | 94.344 | -0.002 | 0.367 | 94.606 |
| | | | (3) $N_A, N_{ab}, N_B, X_A, X_B$ *known* | | | | | |
| *Small* | (105,135) | KL | -0.015 | 0.215 | 94.711 | -0.013 | 0.224 | 94.481 |
| | | EUC | -0.016 | 0.209 | 94.844 | -0.010 | 0.224 | 94.497 |
| | (210,270) | KL | 0.010 | 0.094 | 95.242 | 0.015 | 0.100 | 94.960 |
| | | EUC | 0.011 | 0.094 | 95.279 | 0.015 | 0.101 | 94.914 |
| *Large* | (105,135) | KL | 0.025 | 0.306 | 97.501 | 0.034 | 0.312 | 97.451 |
| | | EUC | 0.031 | 0.278 | 97.731 | 0.038 | 0.301 | 97.542 |
| | (210,270) | KL | -0.004 | 0.121 | 97.883 | -0.001 | 0.134 | 97.663 |
| | | EUC | -0.003 | 0.121 | 97.878 | -0.001 | 0.135 | 97.643 |
| *Medium* | (105,135) | KL | 0.008 | 0.292 | 98.097 | -0.006 | 0.355 | 97.683 |
| | | EUC | 0.010 | 0.276 | 98.197 | -0.008 | 0.356 | 97.676 |
| | (210,270) | KL | 0.001 | 0.128 | 98.120 | 0.003 | 0.173 | 97.465 |
| | | EUC | 0.003 | 0.128 | 98.119 | 0.003 | 0.174 | 97.448 |

Table 5: Length 95% confidence interval, inferior and superior tail error rate, empirical coverage. Linearization and Jackknife variance estimators of the CAL-EUC estimators.

| Linearization/ | | *Single Frame* | | | | *Dual Frame* | | | |
|---|---|---|---|---|---|---|---|---|---|
| Jackknife | $(n_A, n_B)$ | LEN | INF % | SUP % | COV % | LEN | INF % | SUP % | COV % |
| *Sc.1: Small overlap size* | | | | $N_A, N_{ab}, N_B$ *known* | | | | | |
| Lin | (105,135) | 360067 | 1.6 | 2.1 | 96.3 | 365977 | 1.4 | 2.3 | 96.3 |
| Jack | | 337968 | 2.1 | 3.10 | 94.8 | 341008 | 2.2 | 2.6 | 95.2 |
| Lin | (210,270) | 243142 | 2.5 | 1.5 | 96.0 | 249887 | 1.9 | 1.3 | 96.8 |
| Jack | | 233017 | 2.7 | 2.00 | 95.3 | 235151 | 2.8 | 1.8 | 95.4 |
| | | | | $N_A, N_{ab}, N_B, X_A, X_B$ *known* | | | | | |
| Lin | (105,135) | 295610 | 0.1 | 0.6 | 99.3 | 312865 | 0.1 | 1.0 | 98.9 |
| Jack | | 201311 | 2.4 | 3.10 | 94.5 | 203118 | 2.8 | 2.9 | 94.3 |
| Lin | (210,270) | 192396 | 1.3 | 0.6 | 98.1 | 212015 | 0.2 | 0.1 | 99.7 |
| Jack | | 137885 | 3.5 | 1.90 | 94.6 | 139089 | 3.2 | 1.8 | 95.0 |
| *Sc.2: Large overlap size* | | | | $N_A, N_{ab}, N_B$ *known* | | | | | |
| Lin | (105,135) | 458024 | 1.9 | 0.5 | 97.6 | 513298 | 0.3 | 0.4 | 99.3 |
| Jack | | 344233 | 3.6 | 2.9 | 93.5 | 376839 | 3.8 | 2.8 | 93.4 |
| Lin | (210,270) | 292011 | 2.3 | 1.4 | 96.3 | 356127 | 0.5 | 0.1 | 99.4 |
| Jack | | 237054 | 2.3 | 1.9 | 95.8 | 258610 | 2.2 | 2.3 | 95.5 |
| | | | | $N_A, N_{ab}, N_B, X_A, X_B$ *known* | | | | | |
| Lin | (105,135) | 395270 | 0.7 | 1.0 | 98.3 | 441181 | 0.1 | 0.1 | 99.8 |
| Jack | | 207237 | 3.2 | 2.8 | 94.0 | 220257 | 2.8 | 2.7 | 94.5 |
| Lin | (210,270) | 268575 | 1.2 | 0.9 | 97.9 | 303252 | 0.0 | 0.0 | 100. |
| Jack | | 141342 | 2.6 | 2.5 | 94.9 | 149709 | 2.2 | 2.7 | 95.1 |
| *Sc.3: Medium overlap size* | | | | $N_A, N_{ab}, N_B$ *known* | | | | | |
| Lin | (105,135) | 459903 | 1.9 | 1.9 | 96.2 | 485557 | 1.7 | 1.3 | 97.0 |
| Jack | | 394059 | 3.8 | 2.7 | 93.5 | 400930 | 4.1 | 2.2 | 93.7 |
| Lin | (210,270) | 314875 | 1.9 | 1.9 | 96.2 | 336503 | 0.9 | 1.3 | 97.8 |
| Jack | | 276678 | 2.5 | 3.1 | 94.4 | 280932 | 2.6 | 3.4 | 94.0 |
| | | | | $N_A, N_{ab}, N_B, X_A, X_B$ *known* | | | | | |
| Lin | (105,135) | 312413 | 1.6 | 1.6 | 96.8 | 370388 | 0.9 | 0.2 | 98.9 |
| Jack | | 216987 | 2.4 | 3.0 | 94.6 | 220173 | 2.3 | 2.1 | 95.6 |
| Lin | (210,270) | 189223 | 3.2 | 2.8 | 94.0 | 251778 | 0.7 | 0.5 | 98.8 |
| Jack | | 149740 | 2.7 | 2.1 | 95.2 | 151651 | 2.5 | 1.9 | 95.6 |

Table 6: Estimated proportion ($\hat{P}$), lower bound (LB), upper bound (UB) and length (L) of a 95% confidence interval under dual and single frame approach for alternative estimators

| | *Single Frame* | | | | *Dual Frame* | | | |
|---|---|---|---|---|---|---|---|---|
| | $\hat{P}\%$ | LB | UB | L | $\hat{P}\%$ | LB | UB | L |
| HAR | - | - | - | - | 9.03 | 6.10 | 11.95 | 5.85 |
| SF | 11.61 | 8.68 | 14.55 | 5.88 | - | - | - | - |
| PML | - | - | - | - | 11.28 | 7.15 | 15.40 | 8.26 |
| SFRR | 11.25 | 8.81 | 13.70 | 4.89 | - | - | - | - |
| | $N_a, N_{ab}, N_b$ *known* | | | | | | | |
| CAL (*) | 10.97 | 8.68 | 13.27 | 4.60 | 9.49 | 7.08 | 11.90 | 4.82 |
| | $N_a, N_{ab}, N_b, X_A, X_B$ *known* | | | | | | | |
| CAL-EUC | 10.73 | 8.51 | 12.95 | 4.43 | 9.06 | 6.72 | 11.40 | 4.67 |
| CAL-RAK | 10.76 | 8.52 | 12.99 | 4.47 | 9.12 | 6.76 | 11.48 | 4.72 |
| CAL-LOG | 10.76 | 8.53 | 12.99 | 4.46 | 9.11 | 6.75 | 11.47 | 4.71 |
| CAL-KL | 10.71 | 8.47 | 12.95 | 4.48 | 9.22 | 6.78 | 11.66 | 4.88 |

(*) irrespective of the choice of the distance measure