Noname manuscript No.
(will be inserted by the editor)

Population Empirical Likelihood Estimation in Dual Frame Surveys

Maria del Mar Rueda · Maria Giovanna Ranalli · Antonio Arcos · David Molina

Received: date / Accepted: date

Abstract Dual frame surveys are a device to reduce the costs derived from data collection in surveys and improve coverage for the whole target population. Since their introduction, in the 1960's, dual frame surveys have gained much attention and several estimators have been formulated based on a number of different approaches. In this work, we propose new dual frame estimators based on the population empirical likelihood method originally proposed by Chen and Kim (2014) and using both the dual and the single frame approach. The extension of the proposed methodology to more than two frame surveys is also sketched. The performance of the proposed estimators in terms of relative bias and relative mean squared error is tested through simulation experiments. These experiments indicate that the proposed estimators yield better results than other likelihood-based estimators proposed in the literature.

Keywords Multiplicity · Auxiliary information · Multiple frame surveys · Finite population inference

Mathematics Subject Classification (2010) 62D05

Maria del Mar Rueda
Department of Statistics and O. R.
University of Granada
Campus of Fuentenueva, 18110, Granada (Spain) E-mail: mrueda@ugr.es

Maria Giovanna Ranalli
Department of Political Sciences
University of Perugia
Via Pascoli, 06123, Perugia (Italy) E-mail: giovanna.ranalli@unipg.it

Antonio Arcos
Department of Statistics and O. R.
University of Granada
Campus of Fuentenueva, 18110, Granada (Spain) E-mail: arcos@ugr.es

David Molina
Department of Didactics of Mathematics
University of Granada
Campus of Ceuta, 51001, Ceuta (Spain) E-mail: dmolinam@ugr.es

1 Introduction

Classic sampling techniques are based on the concept of a single frame that includes each and every unit of the population. That is, they assume the existence of a complete sampling frame from which the sample is selected. However, in many instances one frame is not enough to guarantee the complete coverage of the whole population. In such cases, the samples drawn from that incomplete frame may suffer from undercoverage, lack of representativeness and, therefore, results may be biased.

Dual frame surveys (Hartley 1962) are a useful tool to face this issue. In a dual frame survey, two sampling frames are available for sampling: each of these frames may be incomplete, but it is assumed that their union covers the entire target population. One independent sample from each frame is selected and, then, data collected from the two samples are combined to produce estimates. Dual frame surveys are useful for reducing cost for given precision constraints, improving coverage and also dealing with elusive or rare populations when a direct sampling frame is not available. When more than two frames are available, multiple frame surveys can be used. Research on dual and multiple frame estimation has been rich since the initial papers by Hartley, and several estimators have appeared in the literature, each derived under a somewhat different approach to estimation. For a good introductory account of dual frame surveys, see the review paper by Lohr (2009); for multiple frame surveys, see Mecatti and Singh (2014).

The growing availability of information coming from census data, administrative registers, and big data provide a wide range of variables, concerning the population of interest, that are eligible to be employed as auxiliary information to increase the efficiency of the estimation procedure. Among these, calibration (Deville and Särndal 1992; Särndal 2007) and regression estimation (Isaki and Fuller 1982) are probably the most popular. Calibration for dual frame surveys has been studied in Ranalli et al. (2016). Calibration and regression estimation have been applied in multiple frame surveys for ordinal responses in Rueda et al. (2018) and in dual frame surveys for categorical responses in Molina et al. (2015). Pseudo-likelihood approaches have also been considered in the literature: Skinner and Rao (1996) proposed a pseudo-maximum likelihood estimator that uses the frame sizes in dual frame surveys to increase the precision of the estimates, while Rao and Wu (2010) formulated a pseudo-empirical likelihood estimator that incorporates auxiliary information in dual and in multiple frame surveys. Recently, Berger and Kabzinska (2019) developed an empirical likelihood multiplicity adjusted estimator that can also handle auxiliary information and that can be applied to a variety of parameters of interest expressed as the unique solution to estimating equations.

Alternative likelihood techniques have been developed to formulate estimators in the classical one frame setting, but they have not been extended to the dual and multiple frame contexts. This is the case of the population empirical likelihood (POEL) estimation approach proposed by Chen and Kim (2014): POEL considers a single empirical likelihood function defined for the finite population. The auxiliary information and the sampling design can be incorporated into the estimation procedure by means of several constraints. Chen and Kim (2014) prove the optimality of this approach with respect to other likelihood-based methods under some unequal sampling designs, such as Poisson sampling. To our knowledge, none of the literature papers on multiple frames estimation addresses the use of POEL as

estimation approach. It is therefore interesting to investigate how to extend this approach to the case of dual and multiple frame surveys and to understand which benefits, if any, it provides with respect to other likelihood approaches. This is the focus of this paper.

The paper is organized as follows. First, we review the pseudo-empirical likelihood estimator of Rao and Wu (2010) to set the framework and the notation (Section 2). Then, we present the proposed estimation approach for the population mean in dual-frame surveys based on the POEL approach in Section 3: we address both the dual and the single frame approach to inference, we discuss extension to other parameters of interest, and we investigate theoretical properties and propose jackknife variance estimation (Sections 3.1–3.5). The extension of this approach to multiple frame surveys is sketched in Section 4. Some simulation experiments are carried out to check the finite size sample properties of the proposed estimators in the dual and in the multiple frame setting (Section 5). Finally, we highlight the most relevant findings and conclusions in Section 6.

2 Pseudo-empirical likelihood estimation in dual frame surveys

We first describe the approach proposed by Rao and Wu (2010). Let U be a finite population composed of N units labeled from 1 to N, $U = \{1, ..., i, ..., N\}$ and let A and B be two overlapping sampling frames, both of them can be incomplete but it is assumed that overall they cover the entire target population U. Let $\mathcal{A}$ and $\mathcal{B}$ be the set of units included in frame A and frame B, respectively, with sizes N_A and N_B. The population of interest may be divided, then, into three disjoint domains: $a = \mathcal{A} \cap \mathcal{B}^c, b = \mathcal{A}^c \cap \mathcal{B}$ and $ab = \mathcal{A} \cap \mathcal{B}$, where ^{c} denotes the complementary of a set. The size of domain a, b, and ab is denoted by N_a, N_b, and N_{ab}, respectively.

Assume we are interested in estimating the population mean of a response variable y, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, where y_i the value of y for the i-th unit of the population. Such mean can be rewritten as

$$\bar{Y} = \frac{N_a}{N}\bar{Y}_a + \frac{N_{ab}}{N}\bar{Y}_{ab} + \frac{N_b}{N}\bar{Y}_b, \tag{1}$$

with $\bar{Y}_a, \bar{Y}_{ab}$, and $\bar{Y}_b$ the population means of the variable Y in domains a, ab, and b, respectively.

Suppose that two samples, s_A and s_B, of sizes n_A and n_B are selected independently from frame A and from frame B, respectively. Let $\pi_i(A) = P(i \in s_A)$ and $\pi_i(B) = P(i \in s_B)$ be the first order inclusion probabilities for units in frame A and in frame B, respectively, and by $d_i(A) = 1/\pi_i(A)$ and $d_i(B) = 1/\pi_i(B)$ the corresponding basic design weights. The units in s_A are such that $s_A = s_a \cup s'_{ab}$, where $s_a = s_A \cap a$ and $s'_{ab} = s_A \cap ab$. Similarly, the units in s_B are such that $s_B = s_b \cup s''_{ab}$, where $s_b = s_B \cap b$ and $s''_{ab} = s_B \cap ab$. Both s'_{ab} and s''_{ab} contain units from the overlap domain ab but the first sample has been selected under the sampling design considered in frame A and the second one is selected through the sampling design considered in frame B.

Rao and Wu (2010) propose the following maximum pseudo-empirical likelihood (PEL) estimator for the mean of a population:

$$\hat{\bar{Y}}_{PEL} = W_a\hat{\bar{Y}}_a + W'_{ab}(\eta)\hat{\bar{Y}}'_{ab} + W''_{ab}(\eta)\hat{\bar{Y}}''_{ab} + W_b\hat{\bar{Y}}_b,$$

where $W_a = N_a/N, W'_{ab}(\eta) = \eta N_{ab}/N, W''_{ab}(\eta) = (1-\eta)N_{ab}/N, W_b = N_b/N$, with $\eta \in (0,1)$ fixed, and $\hat{\bar{Y}}_a = \sum_{i\in s_a} \hat{p}_{a_i} y_i, \hat{\bar{Y}}'_{ab} = \sum_{i\in s'_{ab}} \hat{p}'_{ab_i} y_i, \hat{\bar{Y}}''_{ab} = \sum_{i\in s''_{ab}} \hat{p}''_{ab_i} y_i$ and $\hat{\bar{Y}}_b = \sum_{i\in s_b} \hat{p}_{b_i} y_i$.

The four sets of weights $\hat{p}_a, \hat{p}'_{ab}, \hat{p}''_{ab}$, and $\hat{p}_b$ are such that they maximize the following pseudo empirical likelihood function

$$l_D(p_a, p'_{ab}, p''_{ab}, p_b) = (n_A + n_B) \left\{ W_a \sum_{i\in s_a} \tilde{d}_{a_i} \log(p_{a_i}) + W'_{ab}(\eta) \sum_{i\in s'_{ab}} \tilde{d}'_{ab_i} \log(p'_{ab_i}) \right.$$
$$\left. + W''_{ab}(\eta) \sum_{i\in s''_{ab}} \tilde{d}''_{ab_i} \log(p''_{ab_i}) + W_b \sum_{i\in s_b} \tilde{d}_{b_i} \log(p_{b_i}) \right\}, \tag{2}$$

where $\tilde{d}_{a_i} = d_i(A)/\sum_{i\in s_a} d_i(A), \tilde{d}'_{ab_i} = d_i(A)/\sum_{i\in s'_{ab}} d_i(A), \tilde{d}''_{ab_i} = d_i(B)/\sum_{i\in s''_{ab}} d_i(B)$ and, finally, $\tilde{d}_{b_i} = d_i(B)/\sum_{i\in s_b} d_i(B)$ subject to

$$\sum_{i\in s_a} p_{a_i} = \sum_{i\in s'_{ab}} p'_{ab_i} = \sum_{i\in s''_{ab}} p''_{ab_i} = \sum_{i\in s_b} p_{b_i} = 1$$

and to

$$\sum_{i\in s'_{ab}} p'_{ab_i} y_i = \sum_{j\in s''_{ab}} p''_{ab_j} y_i.$$

The PEL estimator is such that it can incorporate information about auxiliary variables into the estimation process. Assume that a vector of auxiliary variables, $\mathbf{x}_A$, is known for the units in the sample drawn from frame A, so that $\mathbf{x}_{A_i}$ is the value of $\mathbf{x}_A$ for the i-th unit in the frame A. Moreover, the frame population mean $\bar{\mathbf{X}}_A$ is also supposed to be known. This frame-specific information can be incorporated through the constraint

$$\frac{N_a}{N_A} \sum_{i\in s_a} p_{ai} \mathbf{x}_{Ai} + \frac{N_{ab}}{N_A} \sum_{j\in s'_{ab}} p'_{ab_j} \mathbf{x}_{Aj} = \bar{\mathbf{X}}_A.$$

Similarly, a set of constraints can be defined if auxiliary information is known for frame B.

3 Proposed estimators based on population empirical likelihood

Following the POEL approach proposed by Chen and Kim (2014), we can consider the logarithm of the *population level* empirical likelihood as the objective function for the maximization, instead of the *sample level* empirical likelihood used in equation (2). In a dual-frame context, the population empirical log-likelihood can be defined as

$$l(\omega_a, \omega_{ab}, \omega_b) = \frac{N_a}{N} \sum_{i\in a} \log(\omega_{a_i}) + \frac{N_{ab}}{N} \sum_{i\in ab} \log(\omega_{ab_i}) + \frac{N_b}{N} \sum_{i\in b} \log(\omega_{b_i}), \tag{3}$$

with $\omega_{a_i}, \omega_{ab_i}$, and ω_{b_i} such that

$$\sum_{i\in a}\omega_{a_i} = \sum_{i\in ab}\omega_{ab_i} = \sum_{i\in b}\omega_{b_i} = 1. \tag{4}$$

Recall that we denote by $\pi_i(A)$ the first order inclusion probabilities for units in frame A and by $\pi_i(B)$ the first order inclusion probabilities for units in frame B. Let $I_i(A)$ and $I_i(B)$ be the sample selection indicators for frame A and B, respectively. $I_i(A)$ takes value one if unit i is selected in the sample from frame A and takes value zero otherwise. $I_i(B)$ can be defined in a similar way. In the rest of section we assume, as in the case of the pseudo-empirical likelihood approach of Rao and Wu (2010), that the domain sizes N_a, N_{ab}, and N_b are known. However, this assumption can be relaxed as the proposed method can easily accommodate this situation, as it will be shown in the following section.

3.1 POEL – Dual frame approach

Equation (1) can be rewritten as follows

$$\bar{Y} = \frac{N_a}{N}\bar{Y}_a + \eta\frac{N_{ab}}{N}\bar{Y}_{ab} + (1-\eta)\frac{N_{ab}}{N}\bar{Y}_{ab} + \frac{N_b}{N}\bar{Y}_b,$$

with, again, $\eta \in (0,1)$ fixed. Similarly, the population level empirical log-likelihood in (3) can be adapted as follows

$$\begin{aligned} l(\omega_a, \omega_{ab}, \omega_b) = {} & \frac{N_a}{N}\sum_{i\in a}\log(\omega_{a_i}) + \frac{N_{ab}}{N}\eta\sum_{i\in ab}\log(\omega_{ab_i}) \\ & + \frac{N_{ab}}{N}(1-\eta)\sum_{i\in ab}\log(\omega_{ab_i}) + \frac{N_b}{N}\sum_{i\in b}\log(\omega_{b_i}), \end{aligned} \tag{5}$$

and to constraints in (4) we can add

$$\sum_{i\in a}\omega_{a_i}\frac{I_i(A)}{\pi_i(A)} = \sum_{i\in ab}\omega_{ab_i}\frac{I_i(A)}{\pi_i(A)} = \sum_{i\in ab}\omega_{ab_i}\frac{I_i(B)}{\pi_i(B)} = \sum_{i\in b}\omega_{b_i}\frac{I_i(B)}{\pi_i(B)} = 1,$$

to incorporate knowledge of population size of domains a, ab, and b. The previous set of constraints can be rewritten as follows:

$$\sum_{i\in a}\omega_{a_i}\left(\frac{I_i(A)}{\pi_i(A)} - 1\right) = \sum_{i\in ab}\omega_{ab_i}\left(\frac{I_i(A)}{\pi_i(A)} - 1\right) = \sum_{i\in ab}\omega_{ab_i}\left(\frac{I_i(B)}{\pi_i(B)} - 1\right) = \sum_{i\in b}\omega_{b_i}\left(\frac{I_i(B)}{\pi_i(B)} - 1\right) = 0. \tag{6}$$

Moreover, the following constraint

$$\sum_{i\in ab}\omega_{ab_i}\frac{I_i(A)}{\pi_i(A)}y_i = \sum_{i\in ab}\omega_{ab_i}\frac{I_i(B)}{\pi_i(B)}y_i$$

is imposed to assure that estimates for the domain mean of the response variable in the overlap set coincide. Equivalently,

$$\sum_{i\in ab}\omega_{ab_i}\left(\frac{I_i(A)}{\pi_i(A)} - \frac{I_i(B)}{\pi_i(B)}\right)y_i = 0. \tag{7}$$

Finally, assume that population level auxiliary information for frame A is also available. We consider here the case of univariate auxiliary information just for ease of notation. Extension to the multivariate case is straightforward. In particular, we assume that the population mean in frame A of x, $\bar{X}_A = \sum_{i\in\mathcal{A}} x_{A_i}/N_A$ is available. To take this additional information into account, the following constraint

$$\frac{N_a}{N_A}\sum_{i\in a}\omega_{a_i}\frac{I_i(A)}{\pi_i(A)}x_{A_i} + \frac{N_{ab}}{N_A}\sum_{i\in ab}\omega_{ab_i}\frac{I_i(A)}{\pi_i(A)}x_{A_i} = \bar{X}_A \tag{8}$$

is also considered.

We can use the Lagrange multiplier method to compute the three sets of weights $\omega_a, \omega_{ab}, \omega_b$ that maximize (5) subject to (4), (6), (7), and (8). The resulting weights can be expressed as

$$\hat{\omega}_{a_i} = \frac{1}{N_a}\frac{1}{1+\hat{\boldsymbol{\lambda}}^T\mathbf{g}_i}, \quad \hat{\omega}_{ab_i} = \frac{1}{N_{ab}}\frac{1}{1+\hat{\boldsymbol{\lambda}}^T\mathbf{g}_i} \quad \text{and} \quad \hat{\omega}_{b_i} = \frac{1}{N_b}\frac{1}{1+\hat{\boldsymbol{\lambda}}^T\mathbf{g}_i}$$

with

$$\mathbf{g}_i = \left(I_i(a)\left(\frac{I_i(A)}{\pi_i(A)}-1\right), I_i(ab)\left(\frac{I_i(A)}{\pi_i(A)}-1\right), I_i(ab)\left(\frac{I_i(B)}{\pi_i(B)}-1\right), I_i(b)\left(\frac{I_i(B)}{\pi_i(B)}-1\right),\right.$$
$$\left. I_i(ab)\frac{I_i(A)}{\pi_i(A)}y_i - I_i(ab)\frac{I_i(B)}{\pi_i(B)}y_i, (I_i(a)+I_i(ab))\left(\frac{I_i(A)}{\pi_i(A)}x_{A_i}-\bar{X}_A\right)\right)^T$$

where $I_i(u)$ is the indicator variable defined as

$$I_i(u) = \begin{cases} 1 \text{ if } \quad i \in u \\ 0 \text{ otherwise} \end{cases}, \tag{9}$$

and $\hat{\boldsymbol{\lambda}}$ is the solution to $\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\frac{\mathbf{g}_i}{1+\boldsymbol{\lambda}^T\mathbf{g}_i} = \mathbf{0}$.

After computing the three sets of weights, the estimator for the mean can be computed in the following way

$$\hat{\bar{Y}}_{POEL-DF} = \frac{N_a}{N}\hat{\bar{Y}}_a + \eta\frac{N_{ab}}{N}\hat{\bar{Y}}'_{ab} + (1-\eta)\frac{N_{ab}}{N}\hat{\bar{Y}}''_{ab} + \frac{N_b}{N}\hat{\bar{Y}}_b \tag{10}$$

with $\hat{\bar{Y}}_a = \sum_{i\in a}\hat{\omega}_{a_i}\frac{I_i(A)}{\pi_i(A)}y_i, \hat{\bar{Y}}'_{ab} = \sum_{i\in ab}\hat{\omega}_{ab_i}\frac{I_i(A)}{\pi_i(A)}y_i, \hat{\bar{Y}}''_{ab} = \sum_{i\in ab}\hat{\omega}_{ab_i}\frac{I_i(B)}{\pi_i(B)}y_i$ and $\hat{\bar{Y}}_b = \sum_{i\in b}\hat{\omega}_{b_i}\frac{I_i(B)}{\pi_i(B)}y_i$.

In the case in which N_{ab} is not known, the abovementioned procedure can be adapted as follows. Constraints in equation (6) should be replaced by

$$\sum_{i\in a}\omega_{a_i}\frac{I_i(A)}{\pi_i(A)} + \sum_{i\in ab}\omega_{ab_i}\frac{I_i(A)}{\pi_i(A)} = \sum_{i\in ab}\omega_{ab_i}\frac{I_i(B)}{\pi_i(B)} + \sum_{i\in b}\omega_{b_i}\frac{I_i(B)}{\pi_i(B)} = 1.$$

to reflect knowledge of N_A and N_B only, instead of N_a, N_{ab}, and N_b. In addition, the constraint

$$\sum_{i\in ab}\omega_{ab_i}\left(\frac{I_i(A)}{\pi_i(A)} - \frac{I_i(B)}{\pi_i(B)}\right) = 0,$$

could be added to assure that estimates of the overlap domain size coincide. Then the procedure is similar, but with variables $\mathbf{g}_i$ replaced by

$$\mathbf{g}_i = \left((I_i(a) + I_i(ab)) \left(\frac{I_i(A)}{\pi_i(A)} - 1 \right), (I_i(ab) + I_i(b)) \left(\frac{I_i(B)}{\pi_i(B)} - 1 \right), I_i(ab) \left(\frac{I_i(A)}{\pi_i(A)} - \frac{I_i(B)}{\pi_i(B)} \right), \right.$$
$$\left. I_i(ab) \left(\frac{I_i(A)}{\pi_i(A)} - \frac{I_i(B)}{\pi_i(B)} \right) y_i, (I_i(a) + I_i(ab)) \left(\frac{I_i(A)}{\pi_i(A)} x_{A_i} - \bar{X}_A \right) \right).$$

Then, the estimator is given as in (10) with N_a, N_b, and N_{ab} replaced by the corresponding estimates obtained by the final set of weights.

3.2 POEL – Single Frame approach

Another estimator can be computed considering the single frame approach of Bankier (1986) and of Kalton and Anderson (1986). In particular, define and the following single frame inclusion probabilities:

$$\pi_i^\star = \begin{cases} \pi_i(A) & i \in s_a \\ \pi_i(A) + \pi_i(B) & i \in s'_{ab} \cup s''_{ab} \\ \pi_i(B) & i \in s_b \end{cases}. \tag{11}$$

In this case, the likelihood to maximize is

$$l = \sum_{i=1}^{N} \log(\omega_i^\star) \tag{12}$$

subject to the following constraints:

$$\sum_{i=1}^{N} \omega_i^\star = 1, \quad \sum_{i=1}^{N} \omega_i^\star \left(\frac{I_i^c}{\pi_i^\star} - 1 \right) = 0 \quad \text{and} \quad \sum_{i=1}^{N} \omega_i^\star \frac{I_i^c}{\pi_i^\star} x_i = \bar{X},$$

where I_i^c can be seen as a combined indicator which shows whether unit i has been selected in any of the two samples and $\bar{X}$ is the population mean of x. The estimator is computed, then, as

$$\hat{\bar{Y}}_{POEL-SF} = \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_i^\star \frac{I_i^c}{\pi_i^\star} y_i, \tag{13}$$

where

$$\hat{\omega}_i^\star = \frac{1}{N} \frac{1}{1 + \hat{\boldsymbol{\lambda}}^T \mathbf{g}_i},$$

and, in this setting,

$$\mathbf{g}_i = \left(\frac{I_i^c}{\pi_i^\star} - 1, \frac{I_i^c}{\pi_i^\star} x_i - \bar{X} \right).$$

The SF estimator, differently from the proposed DF estimator, does not require knowledge of N_{ab}, but only of N. However, it requires full information (Singh and Mecatti 2011) in order to be computed. That is, one needs identification of frame membership for all units in the population and knowledge of the inclusion probabilities for all the frames from which the unit can be sampled. Full information can be unrealistic in many real surveys, such as those in which unequal probability sampling designs are employed and hinders applicability of this type of estimator.

3.3 Other parameters of interest

The approach has been developed here for estimation of the population mean. However, it can be extended to other parameters of interest of the finite population as long as they can be written as the solution to a set of estimating equations, such as means, quantiles, ratios and generalized linear regression coefficients. That is, consider a parameter θ_0 that is defined by solving $\sum_{i=1}^{N} U(\mathbf{x}_i, y_i; \theta) = 0$ for θ. The population mean is a particular case for which $\theta = \bar{Y}$, with $U(y_i; \bar{Y}) = (y_i - \bar{Y})$.

To encompass this more general situation, we can extend the approach proposed in the original paper by Chen and Kim (2014) to the Dual Frame setting. In particular, we should add the following constraint to the maximization problem described in Section 3.1,

$$\begin{aligned} &\frac{N_a}{N}\sum_{i\in a}\omega_{a_i}\frac{I_i(A)}{\pi_i(A)}U_i(\theta) + \frac{N_{ab}}{N}\eta\sum_{i\in ab}\omega_{ab_i}\frac{I_i(A)}{\pi_i(A)}U_i(\theta) \\ +&\frac{N_{ab}}{N}(1-\eta)\sum_{i\in ab}\omega_{ab_i}\frac{I_i(B)}{\pi_i(B)}U_i(\theta) + \frac{N_b}{N}\sum_{i\in b}\omega_{b_i}\frac{I_i(B)}{\pi_i(B)}U_i(\theta) = 0 \end{aligned} \tag{14}$$

where $U_i(\theta)$ is a shorthand for $U(\mathbf{x}_i, y_i; \theta)$. Then, to solve the complete optimization problem by means of the Lagrange multiplier method, a two-step method is needed. In the first step, the optimal set of weights that maximizes the population level likelihood subject to the constraints discussed in Section 3.1 can be obtained and then used in constraint (14). The same rationale can be applied to the Single Frame setting described in Section 3.2. In this setting, the following constraint should be considered

$$\sum_{i=1}^{N}\omega_i^{\star}\frac{I_i^c}{\pi_i^{\star}}U_i(\theta) = 0.$$

3.4 Asymptotic properties

To show the asymptotic properties of the proposed estimators we adapt and place ourselves in the asymptotic framework of Isaki and Fuller (1982), in which the dual-frame finite population 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, $\{U_N, p_{A_N}(\cdot), p_{B_N}(\cdot)\}$, with $N \to \infty$. We will assume these regularity conditions:

1. N_{A_N} and N_{B_N} tend to infinity and that also n_{A_N} and n_{B_N} tend to infinity as $N \to \infty$;
2. $N_a > 0$ and $N_b > 0$;
3. $n_{A_N}/n_N \to c_1 \in (0,1)$, 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 $N \to \infty$.

This is the same asymptotic framework used in Ranalli et al. (2016) to prove consistency of calibration estimators in dual frame surveys. 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 sequence of designs.

Theorem 1. Under the regularity conditions of Theorem 1 in Chen and Kim (2014) changing π_i *by* $\pi_i^\star$ *the POEL estimator* $\hat{\bar{Y}}_{POEL-SF}$ *has the asymptotic expansion*

$$\hat{\bar{Y}}_{POEL-SF}-\bar{Y} = \frac{1}{N}\sum_{i=1}^{N}\frac{I_i^c}{\pi_i^\star}(y_i-\bar{Y})-B_1^\star(\frac{1}{N}\sum_{i=1}^{N}\frac{I_i^c}{\pi_i^\star}-1)-B_2^\star(\frac{1}{N}\sum_{i=1}^{N}\frac{I_i^c}{\pi_i^\star}(x_i-\bar{X}))+o_p(n^{-1/2})$$

where $(B_1^\star, B_2^\star) = \Omega_1\Omega_2^{-1}$ *with*

$$\Omega_1 = (\frac{1}{N^2}\sum_{i=1}^{N}(\frac{1}{\pi_i^\star}-1)(y_i-\bar{Y}), \frac{1}{N^2}\sum_{i=1}^{N}(y_i-\bar{Y})(x_i-\bar{X}))$$

$$\Omega_2 = \begin{bmatrix} N^{-2}\sum_{i=1}^{N}(\frac{1}{\pi_i^\star}-1) & N^{-2}\sum_{i=1}^{N}(\frac{1}{\pi_i^\star}-1)(x_i-\bar{X}) \\ N^{-2}\sum_{i=1}^{N}(\frac{1}{\pi_i^\star}-1)(x_i-\bar{X}) & N^{-2}\sum_{i=1}^{N}(\frac{1}{\pi_i^\star}-1)(x_i-\bar{X})^2 \end{bmatrix}$$

and has the following asymptotic distribution

$$\frac{\hat{\bar{Y}}_{POEL-SF}-\bar{Y}}{\sqrt{V_\infty}} \to N(0,I),$$

being

$$V_\infty = N^{-2}V(\sum_{i=1}^{N}\frac{I_i^c}{\pi_i^\star}(y_i-\bar{Y}) - B_1^\star(\sum_{i=1}^{N}\frac{I_i^c}{\pi_i^\star}-N) - B_2^\star(\sum_{i=1}^{N}\frac{I_i^c}{\pi_i^\star}(x_i-\bar{X})))$$

Proof. The proof is similar to the proof of Theorem 1 in Chen and Kim (2014) but changing U_i by $(y_i-\bar{Y})$, η_i by $(1, x_i-\bar{X})$, I_i by I_i^c and π_i by $\pi_i^\star$.

3.5 Variance estimation

By Theorem 1 we can obtain a consistent estimator for the variance of POEL-SF estimators, but this estimator is based on asymptotic results. A simple alternative is to use resampling methods. The jackknife approach is a common replication method for variance estimation that can be used in complex surveys for different types of estimators (see e.g. Wolter 2007, for an introduction to jackknife). For the sake of brevity, in this section the proposed estimators are denoted by $\hat{\bar{Y}}_e$, $e = POEL-DF, POEL-SF$. In addition,

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

$$v_J(\hat{\bar{Y}}_e) = V_J^A+V_J^B = \frac{n_A-1}{n_A}\sum_{g\in s_A}(\hat{\bar{Y}}_e^A(g)-\bar{Y}_e^A)^2+\frac{n_B-1}{n_B}\sum_{j\in s_B}(\hat{\bar{Y}}_e^B(j)-\bar{Y}_e^B)^2 \quad (15)$$

where $\hat{\bar{Y}}_e^A(g)$ is the value taken by estimator $\hat{\bar{Y}}_e$ after dropping unit g from s_A and $\bar{Y}_e^A$ is the average of $\hat{\bar{Y}}_e^A(g)$ values. Each value $\hat{\bar{Y}}_e^A(g)$ is computed by excluding unit g from the sample. $\hat{\bar{Y}}_e^B(j)$ and $\bar{Y}_e^B$ are defined similarly. The variance estimator in (15) relies on negligible sampling fractions and can be conservative in case of unequal probability sampling designs such as πps (Wolter 2007).

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

$$v_J^{st}(\hat{\bar{Y}}_e) = V_J^{stA} + V_J^{stB} =$$

$$= \sum_{h=1}^{H} \frac{n_{Ah}-1}{n_{Ah}} \sum_{g \in s_{Ah}} (\hat{\bar{Y}}_e^A(hg) - \bar{Y}_e^{Ah})^2 + \sum_{l=1}^{L} \frac{n_{Bl}-1}{n_{Bl}} \sum_{j \in s_{Bl}} (\hat{\bar{Y}}_e^B(lj) - \bar{Y}_e^{Bl})^2, \quad (16)$$

where $\hat{\bar{Y}}_e^A(hg)$ is the value taken by estimator $\hat{\bar{Y}}_e$ after dropping unit g of stratum h from sample s_{Ah}, $\bar{Y}_e^{Ah}$ is the average of these n_{Ah} values; $\hat{\bar{Y}}_e^B(lj)$ and $\bar{Y}_e^{Bl}$ are defined similarly. In case of a non stratified design in one frame and a stratified design in the other one, previous methods can be combined to obtain the corresponding jackknife estimator of the variance.

4 Extension to more than two frames

In recent years, many papers can be found in the literature that focus on the estimation in cases in which three or more sampling frames are used. Iachan and Dennis (1993) use a three frame survey to reach the homeless population of Washington D.C. metropolitan area. The Canadian Community Health Survey conducted by Statistics Canada (2003) is based on an area frame, a list frame and an RDD frame. Lohr and Rao (2006) formulate the multiple frame extension of some of the estimators originally proposed for the dual frame case, as those proposed by Hartley (1962) and by Fuller and Burmeister (1972). Although the optimal version of these estimators is asymptotically efficient, it is not internally consistent since a different set of weights is used for each response variable. Moreover, it is often unstable in small or moderate samples with more than two frames because the optimal estimated parameters involved in the computation of the estimators are functions of large estimated covariances matrices.

Lohr and Rao (2006) propose a single frame estimator in a multiple frame context. Mecatti (2007) introduces a new approach based on the multiplicity of each unit (i.e. in the number of frames the unit is included in) to propose an estimator which is easy to compute. Singh and Mecatti (2011) generalize this approach and propose the class of Generalized Multiplicity adjusted Horvitz-Thompson design-unbiased estimators. We will focus here on this approach to sketch the proposal of an extension of the POEL approach to more than two frames.

Let $A_1, \ldots, A_q, \ldots, A_Q$ be a collection of $Q \geq 2$ overlapping frames of sizes $N_1, \ldots, N_q, \ldots, N_Q$, all of them can be incomplete but it is assumed that they cover the entire target population U. The population mean can be written as

$$\overline{Y} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{i \in A_q} \frac{y_i}{m_i}, \quad (17)$$

where m_i indicates the number of frames unit i belongs to, i.e. the multiplicity of i. Let s_q be a sample drawn from frame A_q under a particular sampling design and independently for $q = 1, \ldots, Q$, and let $\pi_i(q)$ be the first order inclusion

probabilities under this sampling design. Let n_q be the size of sample s_q and let $s = \cup_q s_q$.

Mecatti (2007) considers a single frame approach and proposes the following single frame multiplicity estimator

$$\hat{\bar{Y}}_M = \frac{1}{N} \sum_{i \in s} \frac{1}{\pi_i(q) m_i} y_i. \tag{18}$$

The single frame multiplicity estimator only requires the knowledge of the multiplicity of each unit, i.e. the number of frames the unit is included, no matter which these frames are. Singh and Mecatti (2011) extend this approach and propose a generalized multiplicity-adjusted methodology for multiple frame estimation. Let $\alpha_i(q)$ be a general multiplicity-adjustment coefficient for every unit i in a given frame U_q with $\sum_q \alpha_i(q) = 1$. A class of design-unbiased estimators is proposed and named Generalized Multiplicity adjusted Horvitz-Thompson (GMHT) estimators:

$$\hat{\bar{Y}}_{GMHT} = \frac{1}{N} \sum_{i \in s} \frac{\alpha_i(q)}{\pi_i(q)} y_i, \tag{19}$$

where the coefficient $\alpha_i(q)$ ensures that y_i is counted once even if unit i is present in more than one frame. The GMHT class encompasses many multiple frame estimators available in the literature. The simple multiplicity-adjusted estimator as given in (18) is the simplest GMHT estimator with the basic choice $\alpha_i(q) = 1/m_i$. The Hartley estimator and the Kalton and Anderson estimator are also GMHT estimators, obtained by making different choices for the multiplicity-adjustment α-coefficient in (19). See Singh and Mecatti (2011) for details on this.

Now we propose a POEL estimator for a collection of $Q > 2$ overlapping frames. Again, for ease of notation, let us consider the case of having auxiliary information on only one variable x_q for each frame q. In particular, let us assume that we know its population mean $\bar{X}_q = \sum_{i \in A_q} x_{qi}/N_q$, for $q = 1, \ldots, Q$. Extension to more auxiliary variables and/or particular cases in which no auxiliary variable is known for some frames can be easily accommodated. In this setting, the population level empirical log-likelihood to maximize can be written as

$$l_{GM} = \sum_{q=1}^{Q} \sum_{i \in A_q} \log(\omega_i(q)) \tag{20}$$

subject to the following constraints:

$$\sum_{i \in A_q} \omega_i(q) = 1, \quad \sum_{i \in A_q} \omega_i(q) \left(\frac{I_i(q)}{\pi_i(q)} - 1 \right) = 0, \quad \text{and} \quad \sum_{i \in A_q} \omega_i(q) \frac{I_i(q)}{\pi_i(q)} x_{qk} = \bar{X}_q$$

for $q = 1, \ldots, Q$, where $I_i(q)$ is an indicator variable which shows whether unit i has been selected in sample s_q. The generalized POEL estimator for multiple frames can then be computed as

$$\hat{\bar{Y}}_{POEL-GM} = \frac{1}{N} \sum_{q=1}^{Q} N_q \sum_{i \in A_q} \hat{\omega}_i(q) \alpha_i(q) \frac{I_i(q)}{\pi_i(q)} y_i, \tag{21}$$

where

$$\hat{\omega}_i(q) = \frac{1}{N_q} \frac{1}{1 + \hat{\boldsymbol{\lambda}}^T \mathbf{g}_i},$$

and, in this setting,

$$\mathbf{g}_i = \left(\left(\frac{I_i(1)}{\pi_i(1)} - 1 \right), \ldots, \left(\frac{I_i(Q)}{\pi_i(Q)} - 1 \right), \frac{I_i(1)}{\pi_i(1)} x_{1k} - \bar{X}_1, \ldots, \frac{I_i(Q)}{\pi_i(Q)} x_{Qk} - \bar{X}_Q \right)^T.$$

The estimate for the set of Lagrange multipliers $\hat{\boldsymbol{\lambda}}$ can be found as the solution to

$$\frac{1}{N} \sum_{q=1}^{Q} \sum_{i \in A_q} \alpha_i(q) \frac{\mathbf{g}_i}{1 + \boldsymbol{\lambda}^T \mathbf{g}_i} = \mathbf{0}.$$

5 Simulation experiments

A comprehensive simulation study has been carried out to check the performance of the proposed estimators. In this study two different experiments are carried out with populations built with two and three frames to compare the dual frame estimators on the one hand, and the generalized estimator for the case $Q = 3$ on the other.

5.1 Dual frame scenarios

An artificial population of size $N = 10000$ is considered. The units of the population have been randomly assigned to one of the three domains ("a", "ab" or "b") with different predetermined domain sizes. We have considered a first scenario with a *small* overlap domain size with N_a=5000, N_b=4000 and, consequently, N_{ab}=1000. The second and the third scenarios have, respectively, *large* and *medium* overlap domain size. The resulting domain sizes in the second scenario are given by N_a=3000, N_b=4000 and N_{ab}=3000, while for the third scenario we have N_a=3000, N_b=2000 and N_{ab}=5000.

Three different sampling designs are used for drawing the sample of each frame: Poisson sampling, simple random sampling and stratified random sampling. Two different combinations of average probabilities of selection (in the case of Poisson sampling) or sample sizes (in the case of simple random sampling and stratified random sampling) are considered. Therefore, the performance of the proposed dual frame estimators has been checked in a total of $3 \times 3 \times 2 = 18$ different scenarios. We focus on checking the behavior of the proposed estimator for various types of sampling designs, various overlap domain sizes and various sample sizes.

The main variable that we have considered is a numeric variable Y with mean 52.78 (sd=9.45). The auxiliary variable X is a numeric variable with mean 52.405 (sd=10.71). The correlation between the two variables is $\rho = 0.60$. Both variables were generated using the `mvrnorm` function from the MASS R-package, (Venables and Ripley 2002). For the Poisson sampling, we use $\pi_i = n z_i / \sum(z_i)$ with $z_i = x_i - a_i$, $a_i \sim N(0, 1)$.

We have computed the two proposed estimators together with the Hartley (Hartley 1962) and the PEL (Rao and Wu 2010) estimators for the purpose of comparison. For each estimator, we compute the percent relative bias $RB\% =$

$E_{MC}[\hat{\bar{Y}} - \bar{Y}]/\bar{Y} * 100$ and the percent relative mean squared error $RMSE\% = E_{MC}[(\hat{\bar{Y}} - \bar{Y})^2]/\bar{Y}^2 * 100$ based on 1000 Monte Carlo simulation runs. We show in Table 1 our results. Relative biases are all negligible, by this providing evidence of successful handling of the dual frame information in the overlap domain. As expected, POEL-SF is always the most efficient estimator as it incorporates more information than the other dual-frame type estimators. Similarly, the Hartley estimator provides the worst performance as it does not include auxiliary information in the estimation procedure. The performance of the proposed POEL-DF is always in line with that of PEL and most of the times it is more efficient. This is particularly true for the stratified design and when the overlap domain size is relatively larger.

5.2 Multiple frame scenarios

The performance of the proposed generalized POEL estimator has been also checked in a multiple frame context. Similarly to the dual frame case, a fictitious population of size $N = 10000$ has been created and its units have been randomly assigned to one of the seven possible domains ("a", "b", "c", "ab", "ac", "bc" or "abc") to simulate a three frame setup. Resulting domain sizes were N_a=1500, N_b=1500, N_c=1000, N_{ab}=2000, N_{ac}=1000, N_{bc}=2000 and, finally, N_{abc}=1000. A different sampling design has been applied in each of the three frames. Then, Poisson sampling has been used to sample from frame A, simple random sampling has been considered to draw samples from frame B, and stratified sampling has been used to sample from frame C. Different combinations of sample sizes were considered:

Sce1: $n_A = 0.030\, N_A$, $n_B = 180$ and $n_C = (25, 30, 35)$;
Sce2: $n_A = 0.065\, N_A$, $n_B = 360$ and $n_C = (50, 60, 75)$;
Sce3: $n_A = 0.130\, N_A$, $n_B = 720$ and $n_C = (100, 120, 150)$.

We use the same y and x variables as in the dual frame simulation study, with $\rho = 0.60$. In addition, we also generated two other pairs of $x - y$ variables with $\rho = 0.70$ and 0.80.

For each simulation study, we compute the single frame multiplicity estimator, $\hat{\bar{Y}}_M$, as a benchmark, and compare it with the proposed estimator $\hat{\bar{Y}}_{POEL-GM}$ and the multiple frame versions of the PEL estimator in Rao and Wu (2010). The two latter estimators exploit the auxiliary information. Comparisons have been carried out using the relative bias and the relative efficiency with respect to the $\hat{\bar{Y}}_M$ estimator, $RE\% = RMSE(\hat{\bar{Y}}_M)/RMSE(\hat{\bar{Y}})$ based on 1000 simulated Monte Carlo replicates. Results for each correlation coefficient may be found in Table 2. Again, relative bias is always negligible for all estimators, meaning that the proposed procedures properly handle multiplicity. The likelihood based methods that use auxiliary information provide much better performances than the basic multiplicity adjusted estimator, and this is more evident as the sample sizes increase (scenarios 1 through 3). In particular, the gain in precision provided here by the proposed estimator that uses a population level likelihood with respect to the PEL that uses a sample level likelihood is more striking. The proposed POEL-GM estimator is always more efficient and this is particularly true when the correlation between the auxiliary variable and the response is stronger.

Table 1 Percent relative mean squared error (RMSE%) and percent relative bias (RB%) of compared estimators.

	RB %			RMSE %		
	Poisson $n_A = 0.0325 \cdot N_A$, $n_B = 0.075 \cdot N_B$					
	Small	*Medium*	*Large*	*Small*	*Medium*	*Large*
$\hat{\bar{Y}}_{Har}$	-0.005	0.032	0.000	0.396	0.370	0.523
$\hat{\bar{Y}}_{PEL}$	-0.018	0.022	0.007	0.282	0.256	0.312
$\hat{\bar{Y}}_{POEL-DF}$	-0.008	0.020	0.017	0.325	0.253	0.300
$\hat{\bar{Y}}_{POEL-SF}$	-0.014	0.020	0.016	0.273	0.245	0.288
	Poisson $n_A = 0.065 \cdot N_A$, $n_B = 0.150 \cdot N_B$					
	Small	*Medium*	*Large*	*Small*	*Medium*	*Large*
$\hat{\bar{Y}}_{Har}$	-0.005	0.023	-0.008	0.187	0.175	0.254
$\hat{\bar{Y}}_{PEL}$	-0.002	0.020	-0.001	0.129	0.121	0.149
$\hat{\bar{Y}}_{POEL-DF}$	0.000	0.019	0.001	0.128	0.120	0.140
$\hat{\bar{Y}}_{POEL-SF}$	0.002	0.015	-0.008	0.125	0.116	0.141
	Stratified $n_{h_A} = (150, 175)$, $n_{h_B} = (50, 60, 75)$					
	Small	*Medium*	*Large*	*Small*	*Medium*	*Large*
$\hat{\bar{Y}}_{Har}$	-0.058	0.112	-0.137	0.662	0.786	0.882
$\hat{\bar{Y}}_{PEL}$	-0.054	0.043	0.030	0.469	0.640	0.561
$\hat{\bar{Y}}_{POEL-DF}$	-0.072	0.033	0.017	0.444	0.607	0.528
$\hat{\bar{Y}}_{POEL-SF}$	-0.093	0.072	0.022	0.420	0.485	0.464
	Stratified $n_{h_A} = (300, 350)$, $n_{h_B} = (100, 120, 150)$					
	Small	*Medium*	*Large*	*Small*	*Medium*	*Large*
$\hat{\bar{Y}}_{Har}$	-0.045	0.077	-0.120	0.345	0.397	0.424
$\hat{\bar{Y}}_{PEL}$	-0.043	-0.017	0.046	0.233	0.312	0.265
$\hat{\bar{Y}}_{POEL-DF}$	-0.064	-0.029	0.032	0.224	0.291	0.244
$\hat{\bar{Y}}_{POEL-SF}$	-0.085	0.022	0.031	0.217	0.225	0.215
	SRS $n_A = 180$, $n_B = 232$					
	Small	*Medium*	*Large*	*Small*	*Medium*	*Large*
$\hat{\bar{Y}}_{Har}$	0.034	-0.006	-0.012	0.811	0.822	1.015
$\hat{\bar{Y}}_{PEL}$	0.018	0.018	0.012	0.617	0.602	0.721
$\hat{\bar{Y}}_{POEL-DF}$	0.018	0.017	0.013	0.619	0.601	0.718
$\hat{\bar{Y}}_{POEL-SF}$	0.022	0.014	0.003	0.605	0.580	0.680
	SRS $n_A = 360$, $n_B = 464$					
	Small	*Medium*	*Large*	*Small*	*Medium*	*Large*
$\hat{\bar{Y}}_{Har}$	-0.001	-0.039	-0.009	0.389	0.400	0.482
$\hat{\bar{Y}}_{PEL}$	0.007	-0.029	0.003	0.290	0.301	0.342
$\hat{\bar{Y}}_{POEL-DF}$	0.006	-0.030	0.005	0.289	0.302	0.339
$\hat{\bar{Y}}_{POEL-SF}$	0.003	-0.029	-0.003	0.284	0.294	0.318

Table 2 Percent relative bias (RB%) and percent relative efficiency (RE%) respect to the M estimator.

	RB %			RE %		
	$Sce1$	$Sce2$	$Sce3$	$Sce1$	$Sce2$	$Sce3$
				$\rho = 0.6$		
$\hat{\bar{Y}}_M$	0.016	0.033	0.037	100.000	100.000	100.000
$\hat{\bar{Y}}_{PEL}$	0.005	0.041	0.031	126.057	127.887	132.348
$\hat{\bar{Y}}_{POEL-GM}$	0.040	0.063	0.061	143.373	142.751	151.281
				$\rho = 0.7$		
$\hat{\bar{Y}}_M$	0.013	0.028	0.036	100.000	100.000	100.000
$\hat{\bar{Y}}_{PEL}$	0.004	0.038	0.027	141.801	145.044	150.063
$\hat{\bar{Y}}_{POEL-GM}$	0.036	0.059	0.056	163.208	162.446	172.044
				$\rho = 0.8$		
$\hat{\bar{Y}}_M$	0.013	0.024	0.034	100.000	100.000	100.000
$\hat{\bar{Y}}_{PEL}$	0.005	0.035	0.024	166.920	171.444	177.047
$\hat{\bar{Y}}_{POEL-GM}$	0.033	0.054	0.050	194.280	192.824	203.693

6 Conclusions

In recent years, multiple-frame surveys have attracted significant attention in survey methodology and applications. The use of more than one frame helps statisticians to obtain more reliable estimates for a finite population, as does the incorporation of available auxiliary population information at different levels. This paper examines an extension of the population empirical likelihood framework proposed by Chen and Kim (2014) to use in estimation from dual-frame surveys. The objective function for the maximization and benchmark constraints are defined and discussed under the single and the dual-frame approaches. The extension of the proposed approach to the multiple-frame setting is also discussed and evaluated in a simulation study.

To define the proposed estimators in a dual frame setting, we have assumed that population sizes N_a, N_b and N_{ab} of the domains are known, as in the case considered for the PEL by Rao and Wu (2010). This situation is common in practice, e.g. when the survey is done by combining landline and cellular phones. However, we have also provided a discussion of how the proposed approach can be adapted when only the frame sizes N_A and N_B are known.

The proposed approach has been applied mainly to the estimation of the population mean of a survey variable. However, it can be extended to a more general class of parameters of interest – i.e. those defined as the solution to a set of estimating equations – as it was proposed in the original population empirical estimation approach by Chen and Kim (2014).

A comprehensive simulation study has been carried out to check the performance of the proposed dual-frame estimators under a number of scenarios with varying sampling designs, sample size, and overlap domain size. Results show that the bias of all proposed estimators is negligible in all scenarios. Additionally, the two proposed dual-frame estimators perform better than the Hartley estimator (that does not include auxiliary information) and the PEL estimator (that uses

the same amount of auxiliary information) in terms of relative mean squared error. This is particularly true for the proposed estimator under the single frame approach. The latter, however, requires full information, i. e., identification of all frame memberships and the inclusion probabilities for each sampled unit for both sampling designs. If samples from frame A and B are both self-weighted, the inclusion probabilities are known but this information is not always available. This fact makes its use restricted in practical applications. An alternative approach in this setting is screening in which units belonging to the overlap are removed from one frame. Nonetheless, this approach can introduce a potential for bias due to nonsampling errors (Kennedy 2007) and, in many cases, it may not be practical or possible to remove list-frame units before sampling.

We have also introduced an extension to more than two frames based on the idea of multiplicity due to Mecatti (2007) and further extended by Singh and Mecatti (2011). The proposed approach allows for the use of frame-specific auxiliary information and is proved to be more efficient that the PEL estimator based on the sample level empirical likelihood, by this providing evidence that the use of the population level empirical likelihood provides better results and is worthy further investigation, in particular in the multiple frame setting.

Acknowledgements This work is partially supported by Ministerio de Economía y Competitividad of Spain (grant MTM2015-63609-R) .

References

Arcos, A., Rueda, M., Martnez-Miranda, M.D. (2005). Using multiparametric auxiliary information at the estimation stage. Statistical Papers 46: 339-358.

Bankier, MD (1986) Estimators based on several stratified samples with applications to multiple frame surveys. J Am Stat Assoc 81:1074–1079.

Berger, YG, Kabzinska, E (2019). Empirical Likelihood Approach for Aligning Information from Multiple Surveys. International Statistical Review.

Chen, S, Kim, JK (2014) Population empirical likelihood for nonparametric inference in survey sampling. Stat Sinica 24:335–355.

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

Fuller, WA, Burmeister, LF (1972) Estimators for samples selected from two overlapping frames. Proceedings of the American Statistical Association, Social Statistics Sections 245–249.

Hartley, HO (1962) Multiple frame surveys. Proceedings of the American Statistical Association, Social Statistics Section.

Iachan, R, Dennis, ML (1993) A multiple frame approach to sampling the homeless and transient population. J Off Stat 9(4):747 – 764.

Isaki, C.T., Fuller, W.A. (1982) Survey Design Under a Regression Superpopulation Model. Journal of the American Statistical Association, 77: 89–96

Kalton, G, Anderson, DW (1986) Sampling rare populations. J R Stat Soc Ser A-G 149:65–82.

Kamgar, S., Meinfelder, F., Mnnich, R. Navvabpour, H. (2018) Estimation within the new integrated system of household surveys in Germany Stat Papers. https://doi.org/10.1007/s00362-018-1023-z

Kennedy, C. (2007) Evaluating the Effects of Screening for Telephone Service in Dual Frame RDD Surveys. Public Opinion Quarterly 71(5):750–771.

Lohr, S, Rao, JNK (2006) Estimation in multiple-frame surveys. J Am Stat Assoc 101(475):1019 – 1030.

Lohr, S (2009) Multiple frame surveys. In D. Pfeffermann and C. R. Rao (Eds). Handbook of Statistics: Vol. 29A. Sample surveys: Design, methods and applications. North Holland, Amsterdam.

Mecatti, F (2007) A single frame multiplicity estimator for multiple frame surveys. Surv Methodol 33:151–158.
Mecatti, F, Singh, A (2014) Estimation in multiple frame surveys: a simplified and unified review using the multiplicity approach. Journal de la Societ Francaise de Statistique 155(4): 55–61.
Molina, D, Rueda, MM, Arcos, A, Ranalli, MG (2015). Multinomial logistic estimation in dual frame surveys. SORT-Statistics and Operations Research Transactions, 39(2), 309–336
Ranalli, MG, Arcos, A, Rueda, MdM, Teodoro, A (2016) Calibration estimation in dual-frame surveys. Stat. Methods Appl. 25-3, 321–349.
Rao, JNK, Skinner, CJ (1996) Estimation in dual frame surveys with complex designs. Proc. of the Survey Method Section, Statistical Society of Canada.
Rao, JNK, Wu, C (2010) Pseudo empirical likelihood inference for multiple frame surveys. J Am Stat Assoc 105:1494–1503.
Rueda, MM, Arcos, A, Molina, D, and Ranalli, MG (2018) Estimation Techniques for Ordinal Data in Multiple Frame Surveys with Complex Sampling Designs. International Statistical Review, 86, 51–67.
Särndal, C.E. (2007). The calibration approach in survey theory and practice. Survey Methodology 33, 99–119.
Singh, AC, Mecatti, F (2011) Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. J Off Stat 27(4):633–650.
Skinner, CJ, Rao, JNK (1996) Estimation in dual frame surveys with complex designs. J Am Stat Assoc 91:349–356.
Venables, WN, Ripley, BD (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York.
Wolter, KM (2007) Introduction to variance estimation, 2nd edn. Springer, Inc., New York.