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

Kernel-based methods for combining information of several frame surveys

Sánchez-Borrego, I. · Arcos, A. · Rueda, M.

Received: date / Accepted: date

Abstract A sample selected from a single sampling frame may not represent adequatly the entire population. Multiple frame surveys are becoming increasingly used and popular among statistical agencies and private organizations, in particular in situations where several sampling frames may provide better coverage or can reduce sampling costs for estimating population quantities of interest. Auxiliary information available at the population level is often categorical in nature, so that incorporating categorical and continuous information can improve the efficiency of the method of estimation. Nonparametric regression methods represent a widely used and flexible estimation approach in the survey context. We propose a kernel regression estimator for dual frame surveys that can handle both continuous and categorical data. This methodology is extended to multiple frame surveys. We derive theoretical properties of the proposed methods and numerical experiments indicate that the proposed estimator perform well in practical settings under different scenarios.

Keywords Kernel regression, nonparametric regression, dual frame survey, multiple frame survey, model-assisted estimation

PACS PACS 62D05 · PACS 62G08

1 Introduction

In classic finite population sampling a basic hypothesis is the availability of a unique and complete list of units forming the target population to be used as a sampling frame. In practice frames that can be used for selecting the samples are generally incomplete or out of date. In some cases a set of two or

Sánchez-Borrego, I., Arcos, A. and Rueda, M.
Department of Statistics and Operational Research. Faculty of Science.
Campus de Fuentenueva, s/n. University of Granada. 18071 Granada, SPAIN.
Tel: +34 958241000 ext. 20067 E-mail: ismasb@ugr.es, arcos@ugr.es, mrueda@ugr.es

more lists is available for survey purposes. 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.

Dual frame estimators were originally introduced by Hartley (1962). Fuller and Burmeister (1972) extended Hartley's dual-frame estimator of population by considering information about the maximum likelihood estimator of the overlap domain population size. Bankier (1986) and Kalton and Anderson (1986) proposed the Single Frame Estimator which treats the dual frame design as a single frame design. Raking ratio or regression estimation can be used to adjust the Single Frame Estimator (Skinner 1991). Skinner and Rao (1986) extended the pseudo maximum likelihood estimator to achieve internal and design-based consistency under complex designs. Rao and Wu (2010) used pseudo empirical likelihood method to include the auxiliary information into the estimation process. Ranalli et al. (2016) used calibration techniques to derive estimators in the dual frame context. The package *Frames2* (Arcos et al. 2015) includes the main estimators in dual frame surveys and also provides interval confidence estimation. Mecatti (2007) proposed a multiplicity estimator for multiple frame surveys, that is insensitive to missclasification. An unified approach for combining information from multiple surveys, by using zero functions as predictors in regression, is given in Singh and Mecatti (2011).

Nonparametric regression methods have been used extensively for estimating the regression function in a wide range of fields. They allow the model to be correctly specified for much larger classes of functions, and have a great potential for application to a wide range of problems. The monograph of Fan and Gijbels (1996) and the chapter in Breidt and Opsomer (2009) explore some application areas for local polynomial regression. Kernel-based methods are well known for their good properties and for their adaptation to different settings.

These methods have been recently introduced into the finite population sampling setting. Kuo (1988) proposed a model-based estimator of the distribution function, Breidt and Opsomer (2000) introduced a model-assisted estimator by incorporating the sampling design to the local linear kernel smoother. Breidt and Opsomer (2000) studied theoretical properties of the local polynomial estimator, showing that it is design consistent and asymptotically design unbiased under some regularity conditions. Nonparametric model calibration has been introduced in Montanari and Ranalli (2005) and used to estimate totals and means also for environmental populations. More recently, Rueda and Sánchez-Borrego (2009) proposed a nonparametric estimator of the total under the model based approach. A review of nonparametric methods in survey sampling is given in Breidt and Opsomer (2009). This work proposes a nonparametric regression approach to inference for multiple-frame surveys. We establish a unified framework for point estimation of finite population parameters, and show that inferences on the parameters of interest makes effective use of the auxiliary population information.

In this article, we investigate theoretical and empirical properties of nonparametric point estimators for the population total when the sample is selected from more than one frame. Multiple frame surveys have been first introduced by Hartley, essentially focused on dual frame case. In order to follow the same methodology as the classic works of multiple frames (Lohr 2009, 2011; Rao and Wu 2010; Metcalf and Scott 2009; etc.) we will start with the case of two frames to later generalize to the case of tree or more frames. Specifically, after introducing the proposed method for dual frame surveys in the second section, we study the asymptotic design-properties of the proposed estimator in the third section. The fourth section considers the problem of selection of the optimal weight. An extension of the nonparametric methodology to multiple frames is proposed in section five and a simulation study is included in section six. Section 7 contains concluding remarks.

2 Nonparametric inference for dual frame surveys

We will use the notation considered in Rao and Wu (2010). Let U denote a finite population with N units, $U = \{1, \ldots, j, \ldots, N\}$ and let A and B be two sampling-frames, both can be incomplete, but it is assumed that together 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, U, may be divided into three mutually exclusive domains, $a = \mathcal{A} \cap \overline{\mathcal{B}}, b = \overline{\mathcal{A}} \cap \mathcal{B}$ and $ab = \mathcal{A} \cap \mathcal{B} = ba$, where the bar symbol denotes complement of a set.

Let N, N_A, N_B, N_a, N_b, N_{ab}, N_{ba} be the number of population units in U, A, B, a, b, ab, ba, respectively.

Let y be a variable of interest in the population, y_j its relative value for the unit j, for $j = 1, \ldots, N$ and $\mathbf{x}_j = (x_{1j}, \ldots, x_{pj})$ the corresponding values of a vector $\mathbf{x}$ containing q categorical and p continuous auxiliary variables, with $q + p = k$. Let $\mathbf{x}^d$ represent the sub-vector of q categorical variables and $\mathbf{x}^c$ the remaining sub-vector of p continuous ones. We use x_t^c for the t-th component of $\mathbf{x}^c$ and x_t^d for the t-th component of $\mathbf{x}^d$, respectively. Our goal is to estimate the finite population total $Y = \sum_{j=1}^{N} y_j$ of the variable of interest y. We can write

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

where $Y_a = \sum_{j \in a} y_j$, $Y_{ab} = \sum_{j \in ab} y_j$, $Y_{ba} = \sum_{j \in ba} y_j$, $Y_b = \sum_{j \in b} y_j$ and η is a fixed constant in $(0, 1)$.

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, using the sampling designs d_A and d_B. Each sampling design d_A and d_B, induces first-order inclusion probabilities π_{A_j} and π_{B_j}, respectively, and second-order inclusion probabilities $\pi_{A_{ij}}$ and $\pi_{B_{ij}}$. The sampling weights are given by $w_{A_j} = 1/\pi_{A_j}$ and $w_{B_j} = 1/\pi_{B_j}$. The sample s_A can be post-stratified 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 $s_{ba} = s_B \cap (ba)$.

We consider a general nonparametric regression model:

$$y_j = m(\mathbf{x}_j) + e_j, \quad j = 1, \ldots, N, \tag{2}$$

where $m(\cdot)$ is the unknown regression function and the e_j, $j = 1, \ldots, N$ are independent and identically distributed with $E_M(e_j) = 0$ and $Var_M(e_j) = \sigma^2$ for $j = 1, \ldots, N$.

We denote E_M and Var_M as the expectation and variance operators under the model and E_{dA}, E_{dB}, V_{dA} and V_{dB} as the expectation and variance operators for sampling designs d_A and d_B respectively.

The *product kernel* for the categorical regressors x_t^d, $t = 1, \ldots, q$ is defined as

$$K_{ij}^d = \prod_{t=1}^{q} l_\lambda(x_{ti}^d - x_{tj}^d), \tag{3}$$

being l_λ a variation of Aitchison-Aitken's kernel function (Aitchison and Aitken 1976), defined as

$$l_\lambda(x_{ti}^d - x_{tj}^d) = \begin{cases} 1 & \text{if} \quad x_{ti}^d = x_{tj}^d \\ \lambda & \text{if} \quad x_{ti}^d \neq x_{tj}^d, \end{cases}, \tag{4}$$

where λ is the smoothing parameter. The parameter λ satisfies $0 \leq \lambda \leq 1$, with $\lambda = 0$ corresponding to an indicator function and $\lambda = 1$ giving equal weights to all values of its argument. For simplicity of notation, we consider a single λ for all variables x_t^d.

For the continuous variables, we use K to denote a symmetric, univariate density function. The product kernel for the mixed data case is defined as

$$\mathcal{K}_{ij} = \left[\prod_{t=1}^{q} l_\lambda(x_{ti}^d - x_{tj}^d)\right] \left[\prod_{t=1}^{p} \frac{1}{h^p} K\left(\frac{x_{ti}^c - x_{tj}^c}{h}\right)\right], \tag{5}$$

being h the smoothing parameter. For simplicity of presentation, we consider a single h for every variable x_t^c, but it can be clearly expanded to a separate h_t for each.

We propose to estimate $m_j = m(\mathbf{x}_j)$, the regression function m at the location $\mathbf{x}_j$, by a design-weighted version of a Nadaraya-Watson (Nadaraya 1964; Watson 1964) type kernel estimator. The estimator $\widehat{m}_j$ is defined as a piecewise function at each domain.

For units j in a, we define

$$\widehat{m}_j^A = \frac{\sum_{i \in s_A} \mathcal{K}_{ij} y_i w_{A_i}}{\sum_{i \in s_A} \mathcal{K}_{ij} w_{A_i} + \delta / N_A^2}, \tag{6}$$

for small appropriate $\delta > 0$. This small order adjustment was also used by Breidt and Opsomer (2000). It ensures the estimator is well-defined for every sample s_A. Without such an adjustment, the estimator may not be defined for some samples. Its effect on the estimator can be made arbitrarily small by choosing δ accordingly.

For domain b we define $\widehat{m}_j^B$ similarly by switching A and B and for the overlap

$$\widehat{m}_j^{ab} = \eta \widehat{m}_j^A + (1-\eta)\widehat{m}_j^B, \tag{7}$$

for $j \in ab$.

We propose the following model-assisted dual frame estimator

$$\begin{aligned}\hat{y}_{\text{np}}^{\text{d}} = &\sum_{j\in a} \widehat{m}_j^A + \sum_{j\in b} \widehat{m}_j^B + \sum_{j\in ab} \widehat{m}_j^{ab} \\ &+ \sum_{j\in s_a} (y_j - \widehat{m}_j^A) w_{A_j} + \sum_{j\in s_b} (y_j - \widehat{m}_j^B) w_{B_j} \\ &+ \eta \sum_{j\in s_{ab}} (y_j - \widehat{m}_j^A) w_{A_j} + (1-\eta) \sum_{j\in s_{ba}} (y_j - \widehat{m}_j^B) w_{B_j}.\end{aligned}$$

Selection of the η-coefficients will be addressed in Section 4. This estimator can be shown to be a particular case of the GMHT estimator (Mecatti and Singh 2014).

3 Properties of the estimator

Theorem 1:

Under assumptions (A1), (A2) and (A4)-(A7) in Breidt and Opsomer (2000) for frames A and B, λ satisfying $0 \leq \lambda \leq 1$, $\hat{y}_{\text{np}}^{\text{d}}$ is asymptotically design unbiased.

Proof: See Appendix

Theorem 2:

Under assumptions (A1)-(A7) in Breidt and Opsomer (2000) for frames A and B, λ satisfying $0 \leq \lambda \leq 1$, the asymptotic variance of the estimator $\hat{y}_{\text{np}}^{\text{d}}$ is given by

$$\begin{aligned}AV\left(\hat{y}_{\text{np}}^{\text{d}}\right) = &\sum_{k,j\in A} (y_j - m_j)(y_k - m_k)\left(w_{A_k} w_{A_j} \pi_{A_{kj}} - 1\right)\eta_{A_k}\eta_{A_j} \\ &+ \sum_{k,j\in B} (y_j - m_j)(y_k - m_k)\left(w_{B_k} w_{B_j} \pi_{B_{kj}} - 1\right)\eta_{B_k}\eta_{B_j},\end{aligned}$$

and the estimator of the variance

$$\begin{aligned}\widehat{V}\left(\hat{y}_{\text{np}}^{\text{d}}\right) = &\sum_{k,j\in s_A} \frac{(y_k - \widehat{m}_k^A)(y_j - \widehat{m}_j^A)\left(w_{A_k} w_{A_j} \pi_{A_{kj}} - 1\right)\eta_{A_k}\eta_{A_j}}{\pi_{A_{kj}}} \\ &+ \sum_{k,j\in s_B} \frac{(y_k - \widehat{m}_k^B)(y_j - \widehat{m}_j^B)\left(w_{B_k} w_{B_j} \pi_{B_{kj}} - 1\right)\eta_{B_k}\eta_{B_j}}{\pi_{B_{kj}}},\end{aligned} \tag{8}$$

is asymptotically design unbiased where

$$\eta_{A_j} = \begin{cases} 1 \text{ if } j \in a \\ \eta \text{ if } j \in ab \end{cases} \tag{9}$$

$$\eta_{B_j} = \begin{cases} 1 \quad \text{if } j \in b \\ 1-\eta \text{ if } j \in ba \end{cases} \tag{10}$$

Proof: See Appendix.

4 Selection of the optimal weight

Selection of parameter η is an important issue in dual frame estimators, because the efficiency of the estimator relies on this value (Lohr 2009). The η-coefficients are a multiplicity adjustment α-coefficients (Mecatti and Singh 2014) that makes estimator $\hat{y}^{\mathrm{d}}_{\mathrm{np}}$ a particular case of the Mecatti's unified estimator.

Skinner and Rao (1996) suggested choosing

$$\eta_{SR} = \frac{N_a N_B V(\hat{N}^B_{ab})}{N_a N_B V(\hat{N}^B_{ab}) + N_b N_A V(\hat{N}^A_{ab})},$$

or alternatively

$$\eta_{SR2} = \frac{V(\hat{N}^B_{ab})}{V(\hat{N}^B_{ab}) + V(\hat{N}^A_{ab})},$$

being $V(\hat{N}^A_{ab})$ and $V(\hat{N}^B_{ab})$ the variances of the sizes of domain ab based on samples s_A and s_B respectively. This value for η does not depend on the y's, thus the resulting estimator use the same weights for the response variable but in practice, N_a, N_b, and variances $V(\hat{N}^A_{ab})$ and $V(\hat{N}^B_{ab})$ are unknown and must be estimated from the data. Thus the use of this value for the weight, leads to internal inconsistency.

Brick et al. (2006) used $\eta = 1/2$ in their study of a dual-frame survey in which frame A was a landline telephone frame and frame B was a cell-phone frame. For this purpose, the value of $\eta = 1/2$ is frequently recommended (Mecatti 2007). Other simple choice for η is $\frac{N_B/n_B}{N_A/n_A+N_B/n_B}$ proposed by Kalton and Anderson (1986) for simple random sampling. These values for η do not depend on the particular y variable being analyzed so that these choices satisfy the requirement of using the same weights for all analyses (Metcalf and Scott 2009).

Hartley (1974) proposed choosing η to minimize the variance of the estimator. Similarly we can derive the optimal value of η for the nonparametric estimator by minimizing $AV\left(\hat{y}^{\mathrm{d}}_{\mathrm{np}}\right)$ respect to η. The optimal asymptotic value is given by

$$\widehat{\eta}_{opt} = \frac{2V_4 + V_2 - V_1}{2V_3 + 2V_4}, \tag{11}$$

with

$$V_1 = \sum_{k\in a;j\in ab} (y_j - m_j)(y_k - m_k)\left(w_{A_k} w_{A_j} \pi_{A_{kj}} - 1\right) + \sum_{k\in ab;j\in a} (y_j - m_j)(y_k - m_k)\left(w_{A_k} w_{A_j} \pi_{A_{kj}} - 1\right),$$

with V_2 analogously defined by switching domain a and frame A for domain b and frame B respectively.

$$V_3 = \sum_{k\in ab;j\in ab} (y_j - m_j)(y_k - m_k)\left(w_{A_k} w_{A_j} \pi_{A_{kj}} - 1\right), \tag{12}$$

being V_4 similarly defined as V_3 but switching frame A for frame B. Nevertheless, since V_1, V_2, V_3 and V_4 are unknown population quantities, they can be estimated by their sample estimates,

$$\widehat{V}_1 = 2 \sum_{k\in s_a;j\in s_{ab}} \frac{(y_j - \widehat{m}_j^A)(y_k - \widehat{m}_k^A)\left(w_{A_k} w_{A_j} \pi_{A_{kj}} - 1\right)}{\pi_{A_{kj}}},$$

and

$$\widehat{V}_3 = \sum_{k\in s_{ab};j\in s_{ab}} \frac{(y_j - \widehat{m}_j^A)(y_k - \widehat{m}_k^A)\left(w_{A_k} w_{A_j} \pi_{A_{kj}} - 1\right)}{\pi_{A_{kj}}}, \tag{13}$$

with $\widehat{V}_2$ and $\widehat{V}_4$ defined as $\widehat{V}_1$ and $\widehat{V}_3$ respectively, with the aboved-mentioned changes.

5 Extension to multiple frames

In recent years, a number of works that focuses on the estimation in cases with three or more sampling frames has arisen. Iachan and Dennis (1993) used 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 a area frame, a list frame and a RDD frame. Lohr and Rao (2006) formulated the multiple frame extension of some of the estimators originally proposed for the dual frame case, as the one proposed by Hartley (1962), Hartley (1974), or 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) also followed the so-called single

frame approach used by Kalton and Anderson (1986) to propose a single frame estimator in a multiple frame context. Mecatti (2007) used 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. Multiplicity is also used by Rao and Wu (2010) to provide an extension of the pseudo empirical likelihood estimator to the case of more than two frames. In 2011, Singh and Mecatti (2011) suggested a class of multiplicity estimators that encompasses all the multiple frames estimators available in the literature by suitably specifying a set of parameters. Recently Rueda et al. (2018) propose statistical techniques for handling ordinal data coming from a multiple frame survey. In this section we propose a nonparametric estimator for the case of more than two frames.

Let U be a finite population composed of N units labeled from 1 to N, $U = \{1, ..., j, ..., N\}$ and 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.

Our goal is to estimate the total population Y, which can be written as

$$Y = \sum_{q=1}^{Q} \sum_{j \in U_q} \frac{y_j}{mu_j} \tag{14}$$

where mu_j indicates the number of frames unit j belongs to, i.e. the multiplicity of j.

Let s_q be a sample drawn from frame A_q under a particular sampling design, independently for $q = 1, \ldots, Q$ and let $\pi_j(q)$ and $\pi_{jk}(q)$ be the first and second order inclusion probabilities under this sampling design, respectively. Let $d_j(q) = 1/\pi_j(q)$ be the sampling weight for units in frame A_q. Let n_q be the size of sample s_q and that $s = \cup_q s_q$.

Mecatti (2007) considered a single frame approach and proposed the following estimator

$$\hat{y}_M = \sum_{j \in s} y_j d_j(q)^M, \tag{15}$$

with $d_j(q)^M = d_j(q)/mu_j$. The previous estimator, often called 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.

In Singh and Mecatti (2011) a generalized multiplicity-adjusted methodology for multiple frame estimation is given. Let $\alpha_j(q)$ be a general multiplicity-adjustment coefficient for every unit j in a given frame U_q with $\sum_q \alpha_j(q) = 1$. A class of design-unbiased estimators is given as Generalized Multiplicity adjusted Horvitz-Thompson (GMHT) class of MF estimators:

$$\hat{y}_{GMHT} = \sum_{j \in s} y_j d_j(q) \alpha_j(q), \tag{16}$$

where the coefficient $\alpha_j(q)$ ensures that y_j is counted once even if unit j is duplicated in more than one frame.

The GMHT class has the potential of encompassing the range of multiple estimators available in the literature. The simple multiplicity-adjusted estimator as given in (15) is the simplest GMHT estimator with the basic choice $\alpha_j(q) = 1/mu_j$. 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 (16) (Singh and Mecatti 2011; Mecatti and Singh 2014). Estimator $\hat{y}^{\mathrm{d}}_{\mathrm{np}}$ can also be shown to be a particular case of estimator (16) in Mecatti and Singh (2014).

For simplicity we consider Mecatti's approach for proposing a nonparametric regression estimator for a collection of $Q \geq 2$ overlapping frames.

For each unit j in the population we can estimate the values m_j in a different form by noting which frame it belongs to. For units j in U_q, we propose to estimate m_j by

$$\widehat{m}_j^q = \frac{\sum_{i \in s_q} \mathcal{K}_{ij} y_i d_i(q)}{\sum_{i \in s_q} \mathcal{K}_{ij} d_i(q) + \delta / N_q^2}, \tag{17}$$

Thus, we propose the multiplicity model-assisted nonparametric estimator as:

$$\hat{y}_{Mnp} = \sum_{q=1}^{Q} \sum_{j \in U_q} \frac{\widehat{m}_j^q}{mu_j} + \sum_{q=1}^{Q} \sum_{j \in s_q} (y_j - \widehat{m}_j^q) \frac{d_j(q)}{mu_j} \tag{18}$$

The estimator (18) depends on the values of the tuning parameters h and λ. We consider allowing their values to be selected by minimizing the multiple-frame cross-validation criterion, defined as

$$CV_{Mnp}(h, \lambda) = \sum_{q=1}^{Q} \frac{1}{n_q} \sum_{i,j \in s_q} \frac{\pi_{ij}(q) - \pi_i(q)\pi_j(q)}{\pi_{ij}(q)} \frac{y_i - \widehat{m}_{-i}}{\pi_i(q)} \frac{y_j - \widehat{m}_{-j}}{\pi_j(q)} \frac{1}{mu_i} \frac{1}{mu_j},$$

where for each $j \in s_q$, $\widehat{m}_{-j}$ is the estimator $\widehat{m}_j$ when the observation (x_j, y_j), $j \in s_q$ is left out.

Although for simplicity of notation, a single λ is introduced for all x_t^d and a single h is defined for all x_t^c, this design-based criterion can also select separate multivariate tuning parameters λ_t and h_t for each explanatory variable.

Theorem 3:

Under assumptions (A1), (A2) and (A4)-(A7) in Breidt and Opsomer (2000) for frames $A_1, ..., A_Q$ and λ satisfying $0 \leq \lambda \leq 1$, $\hat{y}_{Mnp}$ is asymptotically design unbiased.

Theorem 4:

Under assumptions (A1)-(A7) in Breidt and Opsomer (2000) for frames $A_1, ..., A_Q$ and λ satisfying $0 \leq \lambda \leq 1$, the asymptotic variance of the estimator $\hat{y}_{Mnp}$ is given by

$$AV\left(\hat{y}_{Mnp}\right) = \sum_{q=1}^{Q} \sum_{j,k \in U_q} (y_j - m_j)(y_k - m_k)(d_j(q)d_k(q)\pi_{kj}(q) - 1)\frac{1}{mu_k mu_j}. \tag{19}$$

6 Simulation study

In this section, simulations experiments are carried out to illustrate the performance of the multiplicity nonparametric-regression proposed method for a three frame population under different scenarios. We consider a simulated population with size $N = 3500$. Units are randomly assigned to three frames A, B and C according to two different scenarios. In the first scenario (sc1) units are assigned to domains depending on the values taken by a binomial random variable $g_j \sim B(6, 0.3)$. In particular, if $g_j = 0$ then $j \in a$, if $g_j = 1$ then $j \in b$, if $g_j = 2$ then $j \in c$ and for the overlap domains $j \in ab, ac, bc, abc$ if $g_j = 3, 4, 5, 6$, respectively. The resulting sizes of the three frames are N_A=1257, N_B=1733 and N_C=1374 and the overlap domain sizes are N_{ab}=621, N_{bc}=33, N_{ac}=206 and N_{abc}=2. The second scenario (sc2) considers assigning units to domains according to a binomial random variable $g_j \sim B(6, 0.4)$, with probability taken as 0.4 to ensure the overlap domain abc has a few more units. If $g_j = 0$ then $j \in a$, if $g_j = 1$ then $j \in b$, if $g_j = 2$ then $j \in c$ and as before, $j \in ab, ac, bc, abc$ if $g_j = 3, 4, 5, 6$, respectively. The resulting frame sizes in the second scenario are given by N_A=1648, N_B=1756 and N_C=1704 and the overlap domain sizes are N_{ab}=968, N_{bc}=121, N_{ac}=479 and N_{abc}=20. Units in frame B are randomly assigned to five strata as follows: N_h^B=(329, 445, 446, 251, 262) for sc1 and N_h^B=(446, 361, 256, 426, 267) for sc2.

We consider in this simulation study the inclusion of categorical variables in both the underlying population model and in the model-assisted estimator. Let x^d be a binary covariate with $P[x^d = 1] = 0.25$ (used in Sánchez-Borrego et al. 2014). The variable of interest y is generated as a normal distribution $y_j \sim N(5000, 500)$, for $j = 1, \ldots, 3500$. x^c is generated from the values of y as $x_j^c = (y_j - e_j)/0.5$ with $e_j \sim N(500, 500)$, for $j = 1, \ldots, N$. The correlation coefficient with the variable of interest is $\rho = 0.7$ and the population is denoted by LINEAR. The simulations were also performed for different x^c and different values of ρ, but the results were similar and hence are not reported here. The BUMP population is generated by using the mean function $m_1(x) = 1+2(x-0.5)^2+exp(-200(x-0.5)^2)+x^d$ and the CYC population is generated by using the mean function $m_2(x) = 2+2sin(2\pi x)+x^d$, with $x \sim U(0.03, 0.93)$. The errors are generated as independent standard normally random variables and $\sigma = 0.5$. These two regression functions were also used in Breidt and Opsomer (2000).

Samples from frames A and C are selected using simple random sampling without replacement and stratified simple random sampling for frame B. Samples sizes are $n_A = 105$, $n_C = 120$ for the two scenarios and $n_{hB} =$

$(16, 22, 22, 13, 13)$ for the first one and $n_{hB} = (22, 18, 13, 21, 13)$ for the second one.

The proposed method is evaluated in the estimation of the population total. We have applied the above-mentioned multiple-frame cross-validation (CV) criterion for selecting the tuning parameters h and λ. It chooses pairs of h and λ among the set of 15 possible values: $h = 0.1, 0.15, 0.2, 0.3, 0.4$ for the continuous variable and $\lambda = 0, 0.15, 0.3$ for the categorical variable. As the fixed bandwidth $\lambda = 1$ does not take into account the categorical covariate, we have considered small values of λ so that the inclusion of the categorical covariate can improve the efficiency of the proposed estimator. The proposed estimator is computed using the Epanechnikov kernel function (Epanechnikov 1969) for the continuous variable and the Aitchison-Aitken kernel (Aitchison and Aitken 1976) for the categorical variables. For comparison purposes, we also compute Kalton-Anderson (KA, Kalton and Anderson 1976), multiplicity (M, Mecatti 2007) and composite multiplicity (CM, Singh and Mecatti 2011) estimators. We also include in this comparison an extension to three frames of the calibration estimator (CA) given in Ranalli et al. (2016).

We investigated the percent relative bias

$$rb\% = E_{MC}(\hat{Y} - Y)/Y * 100,$$

and the percent relative mean squared error

$$rmse\% = E_{MC}[(\hat{Y} - Y)^2]/Y^2 * 100$$

for each estimator $\hat{Y}$. Simulation results are based on $B = 1000$ samples and E_{MC} denotes the average of Monte Carlo replications.

Table 1 reports results for each scenario. The most efficient estimator in each scenario is denoted in bold.

Table 1 Percent relative bias (rb) and the relative mean squared error ($rmse$) for compared estimators in the three-frames scenarios. Bandwidths h and λ selected by minimizing the multiple-frame cross-validation criterion. Scenario 1 (SC1): Domain size N_a=428, N_b=1077, N_c=1133, N_{ab}=621, N_{bc}=33, N_{ac}=206, N_{abc}=2 ; Samples sizes: n_A= 105, n_B= 86 and n_C= 120. Scenario 2 (SC2): Domain size N_a=181, N_b=647, N_c=1084, N_{ab}=968, N_{bc}=121,N_{ac}=479, N_{abc}=20; Samples sizes: n_A= 105, n_B= 87 and n_C= 120

		Mecatti		KA		CM		CAL		MN	
		RB	RMSE	RB	RMSE	RB	RMSE	RB	RMSE	RB	RMSE
LINEAR	SC1	-0.043	0.004	-0.040	0.004	-0.040	0.004	-0.033	**0.002**	0.032	**0.002**
	SC2	0.034	0.003	0.018	0.004	0.017	0.004	0.020	**0.002**	0.005	**0.002**
BUMP	SC1	0.389	0.096	0.281	0.100	0.276	0.100	0.172	0.016	-0.603	**0.013**
	SC2	-0.428	0.103	-0.664	0.113	-0.674	0.113	-0.029	0.015	-0.560	**0.008**
CYC	SC1	-0.307	0.193	-0.047	0.202	-0.036	0.202	0.055	0.055	0.009	**0.003**
	SC2	0.507	0.181	0.854	0.199	0.86	0.200	0.008	0.052	0.032	**0.003**

We use the survey cross validation method (5) for selecting the bandwidth parameters. The simulations were also performed for different fixed-bandwidth

values, but the results were qualitatively similar and hence are not reported here. The purely-based nonparametric regression estimator is more sensitive to the selection of the bandwith parameter than the design-based nonparametric regression method. Nevertheless, an automated bandwidth selection method is relevant to balance the bias and the variance of the estimates. The survey cross validation criterion we have used seems to work well selecting tuning parameters h and λ and the proposed estimator performs well in all scenarios considered. The use of auxiliary information obviously plays an important role in the estimation process and that clearly translates into the relative bias and efficiency values in table 1. In particular, both the proposed estimator and CA performs better than the other ones that do not take into account the auxiliary information. The proposed estimator is expected to be the preferred estimator, since it does not place any restriction on the relationship between the auxiliary variables and the study variable. The proposed method is close to unbiasedness, as relative biases are less than 1% and it seems to make a better use of the auxiliary information than the CA estimator, as the best results in efficiency for all populations are achieved by the proposed one.

7 Conclusions

In this article, we have presented new estimators to estimate the total of a variable when data are obtained from several frames using nonparametric regression. We have introduced a way to combine estimates from the different frames and considered different estimators based on different level of information.

The first proposed estimator $\hat{y}^{\mathrm{d}}_{\mathrm{np}}$ is based on the same dual frame methodology as in Lohr (2009, 2011), or in Rao and Wu (2010), but it needs full frame level information: the identification of frame membership for every sampled unit and the knowledge of inclusion probability for every frame in which the unit belongs to. It can extended to more than 2 frames, but as noted in Singh and Mecatti (2014), for a collection of $Q \geq 2$ frames available, we have 2^{Q-1} disjoint domains, which may make this extension a laborious and arduous task.

The second proposed estimator $\hat{y}_{Mnp}$, is based on the idea of multiplicity due to Mecatti (2007), is applicable if basic frame level information is available for all sampled units. This information pertain to the selection probability from the sampled frame and the number of frames from which the unit could have been selected but without the frame identification. This approach allows extending to multiple frame in an natural and straightforward way.

Our simulation study shows that the estimator $\hat{y}_{Mnp}$ works well in every scenario and outperforms any other estimator considered, including those that take auxiliary information into account, like the calibration estimator.

We have used the multiplicity estimator due to Mecatti (2007) as a basis for its simplicity, but the estimator $\hat{y}_{Mnp}$ can be extended by using the generalized multiplicity-adjusted methodology introduced by Singh and Mecatti (2011)

in a simple way, changing the weights $1/mu_j$ by $\alpha_j(q)$. The GMHT class has the potential of encompassing the range of MF estimators includes all the known design-unbiased multiple frames estimators and therefore, multiple nonparametric regression estimators can be defined.

8 Appendix

8.1 A.1 Proof of Theorem 1.

We write

$$\begin{aligned}\hat{y}_{\mathrm{np}}^{\mathrm{d}} &= \sum_{j\in a}\widehat{m}_j^A + \sum_{j\in s_a}(y_j-\widehat{m}_j^A)w_{A_j} + \sum_{j\in b}\widehat{m}_j^B \\ &+ \sum_{j\in s_b}(y_j-\widehat{m}_j^B)w_{B_j} + \eta\left(\sum_{j\in ab}\widehat{m}_j^A + \sum_{j\in s_{ab}}(y_j-\widehat{m}_j^A)w_{A_j}\right) \\ &+ (1-\eta)\left(\sum_{j\in ba}\widehat{m}_j^B + \sum_{j\in s_{ba}}(y_j-\widehat{m}_j^B)w_{B_j}\right). \qquad (20)\end{aligned}$$

Let $\hat{y}_{\mathrm{np}}^{a}$ denote the terms $\sum_{j\in a}\widehat{m}_j^A + \sum_{j\in s_a}(y_j-\widehat{m}_j^A)w_{A_j}$. Similarly, $\hat{y}_{\mathrm{np}}^{b}$, $\hat{y}_{\mathrm{np}}^{ab}$ and $\hat{y}_{\mathrm{np}}^{ba}$ denote the corresponding terms on expansion (20).

We write

$$\begin{aligned}\left(\hat{y}_{\mathrm{np}}^{\mathrm{d}}-Y\right) &= \left(\hat{y}_{\mathrm{np}}^{\mathrm{a}}-Y_a\right) + \left(\hat{y}_{\mathrm{np}}^{\mathrm{b}}-Y_b\right) \\ &+ \eta\left(\hat{y}_{\mathrm{np}}^{\mathrm{ab}}-Y_{ab}\right) + (1-\eta)\left(\hat{y}_{\mathrm{np}}^{\mathrm{ba}}-Y_{ba}\right). \qquad (21)\end{aligned}$$

Under (A1), (A2) and (A4)-(A7), λ satisfying $0\le\lambda\le 1$ and taking expectations, Theorem 1 in Breidt and Opsomer (2000) holds for $\left(\hat{y}_{\mathrm{np}}^{\mathrm{a}}-Y_a\right)$ and the same applies to terms $\left(\hat{y}_{\mathrm{np}}^{\mathrm{b}}-Y_b\right)$, $\left(\hat{y}_{\mathrm{np}}^{\mathrm{ab}}-Y_{ab}\right)$ and $\left(\hat{y}_{\mathrm{np}}^{\mathrm{ba}}-Y_{ba}\right)$. Then, the result follows.

8.2 A.2 Proof of Theorem 2.

We write

$$\begin{aligned}\left(\hat{y}_{\mathrm{np}}^{\mathrm{d}}-Y\right) &= \sum_{j\in a}(y_j-\widehat{m}_j^A)(w_{A_j}I_{js_A}-1) + \sum_{j\in ab}(y_j-\widehat{m}_j^A)(w_{A_j}I_{js_A}-1)\eta \\ &+ \sum_{j\in b}(y_j-\widehat{m}_j^B)(w_{B_j}I_{js_B}-1) + \sum_{j\in ba}(y_j-\widehat{m}_j^B)(w_{B_j}I_{js_B}-1)(1-\eta),\end{aligned}$$

where $I_{js_A}=1$ if $j\in s_A$ and $I_{js_A}=0$ otherwise, and $I_{js_B}=1$ if $j\in s_B$ and $I_{js_B}=0$.

The design variance is given by

$$V_d(\hat{y}_{\rm np}^d) = E_{dA}(\sum_{j\in a}(y_j - \widehat{m}_j^A)(w_{A_j}I_{js_A} - 1) + \sum_{j\in ab}(y_j - \widehat{m}_j^A)(w_{A_j}I_{js_A} - 1)\eta)^2$$
$$+ E_{dB}(\sum_{j\in b}(y_j - \widehat{m}_j^B)(w_{B_j}I_{js_B} - 1) + \sum_{j\in ba}(y_j - \widehat{m}_j^B)(w_{B_j}I_{js_B} - 1)(1-\eta))^2$$
$$= E_{dA}(\sum_{j\in A}(y_j - \widehat{m}_j^A)(w_{A_j}I_{js_A} - 1)\eta_{A_j})^2 + E_{dB}(\sum_{j\in B}(y_j - \widehat{m}_j^B)(w_{B_j}I_{js_B} - 1)(\eta_{B_j}))^2,$$

because the sampling designs d_A and d_B are independent. Let

$$c_A = \textstyle\sum_{j\in A}(y_j - m_j)(w_{A_j}I_{js_A} - 1)\eta_{A_j},\ c_B = \sum_{j\in B}(y_j - m_j)(w_{B_j}I_{js_B} - 1)\eta_{B_j},$$
$$t_A = \textstyle\sum_{j\in A}(m_j - \widehat{m}_j^A)(w_{A_j}I_{js_A} - 1)\eta_{A_j} \text{ and } t_B = \sum_{j\in B}(m_j - \widehat{m}_j^B)(w_{B_j}I_{js_B} - 1)\eta_{B_j}.$$

Then

$$E_{dA}\left(\sum_{j\in A}(y_k - \widehat{m}_k^A)(w_{A_j}I_{js_A} - 1)\eta_{A_j}\right)^2 = E_{dA}(c_A + t_A)^2 =$$
$$E_{dA}(c_A^2) + E_{dA}(t_A^2) + 2E_{dA}(t_Ac_A) = E_{dA}(c_A^2) + o(1), \qquad (22)$$

because of lemma 5 in Breidt and Opsomer (2000), $E_{dA}(t_A^2) = o(1)$, so that $E_{dA}(t_Ac_A) \leq (E_{dA}(t_A^2)E_{dA}(c_A^2))^{1/2} = o(1)$.

Similarly $E_{dB}(t_Bc_B) = E_{dB}(c_B^2) + o(1)$.

We have thus that the asymptotic variance of the estimator is given by

$$AV_d(\hat{y}_{\rm np}^{\rm d}) = E_{dA}(c_A^2) + E_{dB}(c_B^2).$$

By using the properties of the Horvitz-Thompson estimator (Horvitz and Thompson 1952) we can deduce

$$E_{dA}(c_A^2) = \sum_{k,j\in A}(y_k - m_k)(y_j - m_j)\left(\sum_{s_A\ni k,j} w_{A_k}w_{A_j}(p_d(s_A) - 1)\eta_{A_k}\eta_{A_j}\right)$$
$$= \sum_{k,j\in A}(y_k - m_k)(y_j - m_j)(w_{A_k}w_{A_j}\pi_{A_{kj}} - 1)\eta_{A_k}\eta_{A_j}. \qquad (23)$$

Using Theorem 3 of Breidt and Opsomer (2000) for the sampling design d_A, we obtain that an unbiased estimator of this variance is given by

$$\sum_{k,j\in s_A}\frac{(y_k - \widehat{m}_k^A)(y_j - \widehat{m}_j^A)\left(w_{A_k}w_{A_j}\pi_{A_{kj}} - 1\right)\eta_{A_k}\eta_{A_j}}{\pi_{A_{kj}}}.$$

A similar expression can be derived for $E_{dB}(c_B^2)$ and then, the result follows.

8.3 A.3 Proof of Theorem 3 and 4.

Proofs of Theorems 3 and 4 are similar to proofs of Theorems 1 and 2; the value of η_{A_j} and η_{B_j} in frames A and B is now assumed by the factors $\frac{1}{mu_j}$ for each frame.

Acknowledgements This research was supported by Ministerio de Economía y Competitividad. Grant number [MTM2015-63609-R] and by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía, Spain).

References

1. Aitchison J, Aitken CGG (1976) Multivariate binary discrimination by the kernel method. Biometrika 63:413-420
2. Arcos A, Molina D, Rueda, M, Ranalli M (2015) Frames2: A Package for Estimation in Dual Frame Surveys. R J 7(1):52–72
3. Bankier M (1986) Estimators based on several stratified samples with applications to multiple frame surveys. J Am Stat Assoc 81:1074–1079
4. Breidt FJ, Opsomer J (2000) Local polynomial regression estimators in survey sampling. Ann Stat 28:1026–1053
5. Breidt FJ, Opsomer J (2009) Nonparametric and semiparametric estimation in complex surveys. In: Rao C, Pfeffermann D (eds) Handbook of Statistics, 29 Part B: Sample Surveys: Theory, Methods and Inference. North Holland, Amsterdam, pp 103-119
6. Brick JM, Dipko S, Presser S, Tucker C, Yuan Y (2006) Nonresponse bias in a dual frame survey of cell and landline numbers. Public Opin Quart 70:780-793
7. Epanechnikov VA (1969) Non-parametric estimation of a multivariate probability density. Theor. Probab. Appl+ 14(1):153-158
8. Fan I, Gijbels I (1996) Local Polynomial Modelling and its Applications. Chapman & Hall, London
9. Fuller W, Burmeister L (1972) Estimators for samples selected from two overlapping frames. Proceedings of social science section of The American Statistical Asociation 101–102
10. Hartley HO (1962) Multiple frame surveys. In: Proceedings of the Social Statistics Section, American Statistical Association 203-206
11. Hartley HO (1974) Multiple frame methodology and selected applications. Sankhya 36:99-118
12. Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Am Stat Assoc 47:663–685
13. Iachan R, Dennis ML (1993) A multiple frame approach to sampling the homeless and transient population. J Off Stat 9:747-764
14. Kalton G, Anderson D (1986) Sampling rare populations. J R Stat Soc Series A (General) 149(1):65–82
15. Kuo L (1988) Classical and prediction approaches to estimating distribution functions from survey data. In ASA Proceedings of the Section on Survey Re-search Methods, American Statistical Association 420:280-285
16. Lohr SL (2011) Alternative survey sample designs: sampling with multiple overlapping frames. Surv Methodol 37(2):197–213
17. Lohr SL (2009) Multiple frame surveys. In: Pfeffermann D, Rao C (eds) Handbook of Statistics, 29 Part A: Sample surveys: Design, Methods and Applications. North Holland, Amsterdam, pp 71–78
18. Lohr S, Rao J (2000) Inference from dual frame surveys. J Am Stat Assoc 95:271–280
19. Lohr S, Rao J (2006) Estimation in multiple-frame surveys. J Am Stat Assoc 101(475):1019–1030
20. Mecatti F, Singh AC (2014) Estimation in Multiple Frame Surveys: A Simplified and Unified Review using the Multiplicity Approach. J-SFdS 155(5):51-69

21. Mecatti F (2007) A single frame multiplicity estimator for multiple frame surveys. Surv Methodol 33(2):151-157
22. Metcalf P, Scott A (2009) Using multiple frames in health surveys. Statist Med 28:1512-1523
23. Montanari GE, Ranalli MG (2005) Nonparametric model calibration estimation in survey sampling. J Am Stat Assoc 100(472):14291442
24. Nadaraya EA (1964) On estimating regression. Theor Probab Appl+ 9(1):141-142
25. Ranalli MG, Arcos A, Rueda M, Teodoro A (2016) Calibration estimation in dual-frame surveys. Stat Method Appl 25(3):321-349
26. Rao J, Wu C (2010) Pseudoempirical likelihood inference for multiple frame surveys. J Am Stat Assoc 105(492):1494–1503
27. Rueda M, Sánchez-Borrego I (2009) A predictive estimator of finite population mean using nonparametric regression. Computation Stat 24:1-14
28. Rueda M, Arcos A, Molina D, Ranalli M (2017) Estimation techniques for ordinal data in multiple frame surveys with complex sampling designs, International Statistical Review (in press) https://doi.org/10.1111/insr.12218
29. Sánchez-Borrego I, Opsomer J, Rueda M, Arcos A (2014) Nonparametric estimation with mixed data types in survey sampling. Rev Mat Complut 27(2):685-700
30. 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
31. Skinner C (1991) On the eficiency of raking ratio estimation for multiple frame surveys. J Am Stat Assoc 86(415):779–784
32. Skinner C, Rao J (1996) Estimation in dual frame surveys with complex designs. J Am Stat Assoc 91(433):349–356
33. Watson, GS (1964) Smooth regression analysis. Sankhya Series A 26(4):359-372