SORT 39 (2) July-December 2015, 309-336

Multinomial logistic estimation in dual frame surveys

David Molina[1], Maria del Mar Rueda[1], Antonio Arcos[1]
and Maria Giovanna Ranalli[2]

Abstract

We consider estimation techniques from dual frame surveys in the case of estimation of proportions when the variable of interest has multinomial outcomes. We propose to describe the joint distribution of the class indicators by a multinomial logistic model. Logistic generalized regression estimators and model calibration estimators are introduced for class frequencies in a population. Theoretical asymptotic properties of the proposed estimators are shown and discussed. Monte Carlo experiments are also carried out to compare the efficiency of the proposed procedures for finite size samples and in the presence of different sets of auxiliary variables. The simulation studies indicate that the multinomial logistic formulation yields better results than the classical estimators that implicitly assume individual linear models for the variables. The proposed methods are also applied in an attitude survey.

MSC: 62D05

Keywords: Finite population, survey sampling, auxiliary information, model assisted inference, calibration.

1. Introduction

Sampling theory for finite populations usually assumes the existence of one sampling frame containing all population units. Then, a probability sample is drawn according to a sampling design and information collected is used for estimation and inference purposes. To ensure quality of the results obtained, the sampling frame must contain every single unit of population of interest (that is, it must be complete) and it must be updated as well. Otherwise, estimates could be affected by a serious bias due to the non-representativeness of the frame and, therefore, of the selected sample. Unfortunately,

[1] Department of Statistics and Operational Research, University of Granada, Spain. dmolinam@ugr.es, mrueda@ugr.es, arcos@ugr.es

[2] Department of Political Sciences, University of Perugia, Italy. giovanna@stat.unipg.it

Received: September 2014
Accepted: October 2015

this is not an easy task: populations are constantly changing, with new units entering and exiting the population frequently, so getting a good sampling frame can be difficult.

The dual frame approach tries to solve the aforementioned problems. This approach assumes that two frames are available for sampling and that, overall, they cover the entire target population. A sample is selected from each frame using a, possibly different, sampling design. Much attention has been devoted to the introduction of different ways of combining estimates coming from the different frames – see the seminal papers by Hartley (1962), Fuller and Burmeister (1972), Bankier (1986) and Kalton and Anderson (1986). However, these techniques were originally proposed to estimate means and totals of quantitative variables, and although their extension to the estimation of proportions in multinomial response variables is possible, it requires further investigation. Questionnaire items with multinomial outcomes are quite common in public opinion research, marketing research, and official surveys: estimating the proportion of voters in favour of each political party, based on a political opinion survey, is just one practical example of this procedure. Items where respondents must select one in a series of options can be modeled by a multinomial distribution. Lehtonen and Veijanen (1998) present estimators for a proportion which use logistic regression.

This paper focuses on the estimation of proportions for multinomial response variables when data come from two sampling frames. The proposed approach is motivated by a study on immigration. After describing the survey of opinions and attitudes of the Andalusian population regarding immigration, in Section 2, alternative estimators for the proportions are proposed following different approaches and their main theoretical properties are studied. A simulation study is also carried out to study their finite size sample properties. The results from the application to this dual frame attitude survey are then presented in Section 9.

2. Study background: the 2013 survey on opinions and attitudes of the Andalusian population regarding immigration

The 2013 survey on opinions and attitudes of the Andalusian population regarding immigration (OPIA) is a population-based survey conducted by the *Instituto de Estudios Sociales Avanzados* (IESA), a public scientific research institute for social sciences. The aim of the survey is to reflect the opinion of the Andalusian population with regard to various aspects of immigration and refugee policies in Spain and towards immigrants as a group. This survey is based on telephone interviews on a sample of adults drawn from both landline and mobile phone frames. Taking into account the time and budget available, 2402 interviews were performed by professional interviewers. The number of interviews to be conducted via landline and via mobile phone was determined by calculating the optimal proportion (in the sense of minimum variance) for each frame, taking into account costs and the percentage of possession of each type of device (following Hartley (1962)). As a result, final sample sizes were 1919 for landline and 483

Table 1: *Sample sizes for the OPIA survey. Landline and Mobile in the columns refer to the frame the interview comes from, while in the rows, they refer to the domain in which the units actually reside (type of user).*

Domain	Landline Sample	Mobile Sample	Total
Both	1 727	237	1 964
Mobile		246	246
Landline	192		192
Total	1 919	483	2 402

for mobile. Interviews were carried out by the Statistics and Surveys sections of IESA from April, 22 to May, 13, 2013, using Computer Assisted Telephone Interviewing (CATI) data input techniques. Sample sizes are reported in Table 1. The landline sample was also stratified by provinces in the region of Andalusia, as shown in Table 2. Cell-phone interviews were carried out with no control over the distribution by provinces owing to the difficulty of determining the location of this type of telephone. Hence, more interviews were performed in the most populated provinces than in the less populated ones.

Table 2: *Stratification in land-phone sample.*

Province	Almería	Cádiz	Córdoba	Granada	Huelva	Jaén	Málaga	Sevilla
Population(*)	353 787	767 370	508 258	558 087	308 941	423 548	872 011	1 190 918
Sample	262	210	252	256	275	263	207	194

(*) Those estimates can be found on the INE website: http://www.ine.es/

At the time of data collection, frame sizes of landline and mobile were 4 982 920 and 5 707 655, respectively, and the total population size was 6 350 916 (source ICT-H 2012, Survey on the Equipment and Use of Information and Communication Technologies in Households, INE, National Statistical Institute, Spain). Auxiliary information about the user's sex and age is also available from the ICT-H 2012 survey. The total number of individuals in each domain (landline, mobile and both users) for every possible combination of values of the auxiliary variables is therefore known. The information about these auxiliary variables is displayed in Table 3.

One of the most important response variables in this study is related to the "attitude towards immigration". The variable is the answer to the following question: *And in relation to the number of immigrants currently living in Andalusia, do you think there are ...?: Too many, A reasonable number, Too few, No reply*. In the following sections we review approaches available in the literature to address the issue of estimating the distribution of a multiple choice type of variable in the population using a dual frame survey. We then illustrate our proposal to fully account for the nature of the response variable and the auxiliary information available.

Table 3: *Population data for variables* sex *and* age.

	Both	Landline	Mobile	Total
		Males		
18 - 29	428 750	0	188 172	616 922
30 - 44	724 435	4 259	298 416	1027 110
45 - 59	603 338	59 385	135 981	798 704
≥ 60	396 626	206 410	94 729	697 765
		Females		
18 - 29	480 151	0	115 472	595 623
30 - 44	658 984	17 673	289 106	965 763
45 - 59	601 478	39 362	141 553	782 393
≥ 60	445 897	316 172	104 567	866 636

(*) Source: Survey of Information Technologies in Households (INE)

3. Existing approaches to estimation of class frequencies in dual frame surveys

We employ the notation considered in Rao and Wu (2010). Let U denote a finite population with N units, $U = \{1,\ldots,k,\ldots,N\}$ and let A and B be two sampling-frames. Let $\mathscr{A}$ be the set of population units in frame A and $\mathscr{B}$ the set of population units in frame B. The population of interest, U, may be divided into three mutually exclusive domains, $a = \mathscr{A} \cap \mathscr{B}^c, b = \mathscr{A}^c \cap \mathscr{B}$ and $ab = \mathscr{A} \cap \mathscr{B}$. 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 $ba = \mathscr{B} \cap \mathscr{A}$, which is identical to $ab = \mathscr{A} \cap \mathscr{B}$, 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 U, A, B, a, b, ab, ba, respectively. We assume that N_A, N_B and N_{ab} are known, so the population size $N = N_A + N_B - N_{ab}$ is also known. This is also the situation in our motivating dataset.

We consider the estimation of class frequencies of a discrete response variable. Assume that we collect data from respondents who provide a single choice from a list of alternatives. We code these alternatives $1,2,\ldots,m$. Therefore, consider a discrete m-valued survey variable y. The objective is to estimate the frequency distribution of y in the population U. To estimate this frequency distribution, we define a class of indicators z_i $(i = 1,\ldots,m)$ such that, for each unit $k \in U$, $z_{ki} = 1$ if $y_k = i$ and $z_{ki} = 0$ otherwise. Our problem thus, is to estimate the proportions $P_i = N^{-1}\sum_{k\in U} z_{ki}$, for $i = 1,2,\ldots,m$. These proportions are such that

$$P_i = N^{-1}(Z_{ai} + \eta Z_{abi} + (1-\eta)Z_{bai} + Z_{bi}), \tag{1}$$

where $0 \leq \eta \leq 1$ and $Z_{ai} = \sum_{k\in a} z_{ki}$, $Z_{abi} = \sum_{k\in ab} z_{ki}$, $Z_{bai} = \sum_{k\in ba} z_{ki}$ and $Z_{bi} = \sum_{k\in b} z_{ki}$.

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

π_{Ak} and π_{Bk}, respectively, and sampling weights $d_{Ak} = 1/\pi_{Ak}$ and $d_{Bk} = 1/\pi_{Bk}$. 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 $s_{ba} = s_B \cap (ba)$. 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. Then, assuming that duplicated units (i.e. $s_A \cap s_B$) cannot be identified and that this event has a negligible chance to happen, we let $s = s_A \cup s_B$. Note that this is a reasonable assumption in the OPIA survey at hand.

The Hartley (1962) estimator of P_i, for $i = 1, 2, \ldots, m$, is given by

$$\hat{P}_{Hi}(\eta) = N^{-1}(\hat{Z}_{ai} + \eta\hat{Z}_{abi} + (1-\eta)\hat{Z}_{bai} + \hat{Z}_{bi}), \tag{2}$$

where $\hat{Z}_{ai} = \sum_{k \in s_a} d_{Ak} z_{ki}$ is the expansion estimator for the population count of category i in domain a and similarly for the other domains. If we let

$$d_k^\circ = \begin{cases} d_{Ak} & \text{if } k \in s_a \\ \eta d_{Ak} & \text{if } k \in s_{ab} \\ (1-\eta) d_{Bk} & \text{if } k \in s_{ba} \\ d_{Bk} & \text{if } k \in s_b \end{cases}, \tag{3}$$

then $\hat{P}_{Hi}(\eta) = N^{-1}(\sum_{k \in s_A} d_k^\circ z_{ki} + \sum_{k \in s_B} d_k^\circ z_{ki}) = N^{-1}(\sum_{k \in s} d_k^\circ z_{ki})$. Since the population count in each domain is estimated by its expansion estimator, $\hat{P}_{Hi}(\eta)$ is an unbiased estimator of P_i for a given η.

Fuller and Burmeister (1972) proposed modifying Hartley's estimator by incorporating additional information regarding estimation of the overlap domain. The resulting estimator is:

$$\hat{P}_{FBi}(\beta_1, \beta_2) = N^{-1}(\hat{Z}_{ai} + \beta_1\hat{Z}_{abi} + (1-\beta_1)\hat{Z}_{bai} + \hat{Z}_{bi} + \beta_2(\hat{N}_{ab} - \hat{N}_{ba})) \tag{4}$$

where $\hat{N}_{ab} = \sum_{k \in s_{ab}} d_{Ak}$ and $\hat{N}_{ba} = \sum_{k \in s_{ba}} d_{Bk}$. Coefficients β_1 and β_2 are selected to minimize $V(\hat{P}_{FBi}(\beta_1, \beta_2))$. In this case, and as with Hartley's estimator, a new set of weights must be calculated for each response variable. This leads to possible inconsistencies among the estimated proportions, which is particularly relevant when dealing with multinomial outcomes. In addition, optimal values depend on covariances among Horvitz-Thompson estimators, which may be difficult to compute in practice and, finally, it is also possible to obtain values of β_1 outside the range $[0, 1]$.

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:

$$\hat{P}_{PMLi}(\theta) = N^{-1}\left(\frac{N_A - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a}\hat{Z}_{ai} + \frac{\hat{N}_{ab}^{PML}(\theta)}{\hat{N}_{ab}(\theta)}\hat{Z}_{abi}(\theta) + \frac{N_B - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b}\hat{Z}_{bi}\right) \quad (5)$$

where $\hat{Z}_{abi}(\theta) = \theta\hat{Z}_{abi} + (1-\theta)\hat{Z}_{bai}, \hat{N}_{ab}(\theta) = \theta\hat{N}_{ab} + (1-\theta)\hat{N}_{ba}$ and $\hat{N}_{ab}^{PML}(\theta)$ is the smallest root of the quadratic equation

$[\theta/N_B + (1-\theta)/N_A]x^2 - \left[1 + \theta\hat{N}_{ab}/N_B + (1-\theta)\hat{N}_{ba}/N_A\right]x + \hat{N}_{ab} = 0.$

Recently, Rao and Wu (2010) extended the Pseudo-Empirical-Likelihood approach (PEL) proposed by Wu and Rao (2006) from one-frame surveys to dual-frame surveys following a stratification approach. In particular,

$$\hat{P}_{PELi}(\theta) = (N_a/N)\hat{\bar{Z}}_{aip} + \theta(N_{ab}/N)\hat{\bar{Z}}_{abip} + (1-\theta)(N_{ba}/N)\hat{\bar{Z}}_{baip} + (N_b/N)\hat{\bar{Z}}_{bip}, \quad (6)$$

where $\theta \in (0,1)$ is a fixed constant to be specified and $\hat{\bar{Z}}_{aip} = \sum_{k\in s_a}\hat{p}_{ak}z_{ki}$, $\hat{\bar{Z}}_{bip} = \sum_{k\in s_b}\hat{p}_{bk}z_{ki}$ and $\hat{\bar{Z}}_{abip} = \sum_{k\in s_{ab}}\hat{p}_{abk}z_{ki} = \hat{\bar{Z}}_{baip}$. The p-weights maximize the pseudo empirical likelihood and verify $\sum_{k\in s_a}\hat{p}_{ak} = 1, \sum_{k\in s_{ab}}\hat{p}_{abk} = 1, \sum_{k\in s_{ba}}\hat{p}_{bak} = 1, \sum_{k\in s_b}\hat{p}_{bk} = 1$, and the additional constraint induced by the common domain mean $\hat{\bar{Z}}_{abip} = \hat{\bar{Z}}_{baip}$ (see Rao and Wu (2010) for more details). Note that (6) can be rewritten as:

$$\hat{P}_{PELi} = (N_a/N)\hat{\bar{Z}}_{aip} + (N_{ab}/N)\hat{\bar{Z}}_{abip} + (N_b/N)\hat{\bar{Z}}_{bip}, \quad (7)$$

so the estimator does not depend on explicitly on θ and its value only affects the estimator $\hat{\bar{Z}}_{abip}$ for the population mean of the overlapping domain.

Ranalli et al. (2015) used calibration procedures for estimation from dual frame sampling assuming that some kind of auxiliary information is available. For example, assuming that there are p auxiliary variables, $\boldsymbol{x}_k = (x_{1k}, ..., x_{pk})$ is the value taken by such auxiliary variables on unit k. It is assumed that the vector of population totals of the auxiliary variables, $\boldsymbol{t}_x = \sum_{k\in U}\boldsymbol{x}_k$ is also known. In this context, the dual frame calibration estimator can be defined as follows,

$$\hat{P}_{CalDFi} = N^{-1}\left(\sum_{k\in s} d_k^{\mathrm{DF}} z_{ki}\right) \quad (8)$$

where weights d_k^{DF} are chosen to be as close as possible to basic design weights and, at the same time, satisfy benchmark constraints on the auxiliary variables, i.e. they are such that

$$\min_{d_k^{\mathrm{DF}}} \sum_{k\in s} G(d_k^{\mathrm{DF}}, d_k^{\circ}), \quad \text{subject to} \quad \sum_{k\in s} d_k^{\mathrm{DF}}\boldsymbol{x}_k = \boldsymbol{t}_x,$$

with $G(\cdot,\cdot)$ a given distance measure.

When inclusion probabilities in domain *ab* are known for both frames, and not just for the frame from which the unit is selected, *single-frame* methods (Bankier (1986), Kalton and Anderson (1986)), which combine the observations into a single dataset and adjust the weights in the intersection domain for multiplicity, can also be used. To adjust for multiplicity, the weights are defined as follows for all units in frame A and in frame B,

$$\tilde{d}_k = \begin{cases} d_{Ak} & \text{if } k \in a \\ (1/d_{Ak} + 1/d_{Bk})^{-1} & \text{if } k \in ab \\ d_{Bk} & \text{if } k \in b \end{cases} .$$

In this context, BKA single frame estimator (Bankier (1986) and Kalton and Anderson (1986)) is given by

$$\hat{P}_{BKAi} = N^{-1} \left(\sum_{k \in s_A} \tilde{d}_k z_{ki} + \sum_{k \in s_B} \tilde{d}_k z_{ki} \right) = N^{-1} \left(\sum_{k \in s} \tilde{d}_k z_{ki} \right). \tag{9}$$

Single frame weights are the same for all response variables, and so estimators are internally consistent.

A calibration estimator under the *single-frame* approach can be defined as follows:

$$\hat{P}_{CalSFi} = N^{-1} \left(\sum_{k \in s} d_k^{SF} z_{ki} \right) \tag{10}$$

with weights d_k^{SF} verifying that $\min \sum_{k \in s} G(d_k^{SF}, \tilde{d}_k)$ subject to $\sum_{k \in s} d_k^{SF} \boldsymbol{x}_k = \boldsymbol{t}_x$.

The single-frame approach requires the knowledge of the design weight of a unit for both frames, not just for the one in which the unit was selected. Given this information, multiplicity can be adjusted for using sampling weights only. Therefore, unlike the dual frame methods, they do not require calculation of η. Single-frame estimators are usually more efficient than dual-frame estimators, and this can be explained by the extra-information they incorporate in the estimation process. The estimators presented in this Section can be computed using the R-package `Frames2` (Arcos et al., 2015).

4. Estimation of class frequencies using multinomial logistic regression

Auxiliary information is often available in survey sampling. This information, which may come from past censuses or from other administrative sources, can be used to obtain more accurate estimators. Then, other than the values of the variables of interest and of

the auxiliary variables for $k \in s$, assume we also know the distribution or at least some summary statistics of the auxiliary variables in the population. We consider that the population under study $\boldsymbol{y} = (y_1, ..., y_N)^\top$ is the determination of a set of super-population random variables $\boldsymbol{Y} = (Y_1, ..., Y_N)^\top$ s.t.

$$\mu_{ki} = P(Y_k = i|\boldsymbol{x}_k) = E(Z_{ki}|\boldsymbol{x}_k) = \frac{\exp(\boldsymbol{x}_k^\top \boldsymbol{\beta}_i)}{\sum_{r=1,...,m} \exp(\boldsymbol{x}_k^\top \boldsymbol{\beta}_r)}, \qquad i = 1, \ldots, m,$$

that is, we use the multinomial logistic model to relate y and $\boldsymbol{x}$. Let $\boldsymbol{\beta}$ be the parameter vector $(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_m^\top)^\top$. In the following sections we introduce new estimators for the population proportions P_i. To this end, as a first step, we need to consider estimation of the superpopulation parameter $\boldsymbol{\beta}$ using the sample s.

4.1. Case I: The same set of auxiliary variables is available for all population units

Suppose that for each unit in the population we have information about one vector of auxiliary variables $\boldsymbol{x}$. In this case, for each unit $k \in U$ we know the value of $\boldsymbol{x}_k$. In addition, for each unit $k \in s$, we observe the value of the main variable y_k and we denote by $(z_{k1}, z_{k2}, ..., z_{km})$ the multinomial trial observed for this unit k.

We can estimate $\boldsymbol{\beta}$ by maximizing the π-weighted log-likelihood (Godambe and Thompson (1986), Särndal et al. (1992)) given by

$$\ell_{d^\circ}(\boldsymbol{\beta}) = \sum_{i=1,...,m} \left(\sum_{k \in s_A} d_k^\circ z_{ki} \ln \mu_{ki} + \sum_{k \in s_B} d_k^\circ z_{ki} \ln \mu_{ki} \right). \tag{11}$$

This approach is usually motivated by first defining a census-level parameter $\boldsymbol{\beta}_U$, obtained by maximizing the likelihood over all units in the population, i.e. $\ell_U(\boldsymbol{\beta}) = \sum_{i=1,...,m} \sum_{k \in U} z_{ki} \ln \mu_{ki}$. Then, $\widehat{\boldsymbol{\beta}}^\circ$ obtained using the the π-weighted likelihood (11) is its design based estimate. Computing $\widehat{\boldsymbol{\beta}}^\circ$ usually requires numerical procedures, and Fisher scoring or Newton-Raphson often work rather well. Most statistical packages include a multinomial logit procedure that can handle weights.

Given the estimate $\widehat{\boldsymbol{\beta}}^\circ$ of $\boldsymbol{\beta}$, we consider the following auxiliary variable

$$p_{ki}^\circ = \widehat{\mu}_{ki}^\circ = \frac{\exp(\boldsymbol{x}_k^\top \widehat{\boldsymbol{\beta}}_i^\circ)}{\sum_{r=1,...,m} \exp(\boldsymbol{x}_k^\top \widehat{\boldsymbol{\beta}}_r^\circ)}. \tag{12}$$

Please note that these p values are different from those involved in the definition of estimator (6). Since the vector $\boldsymbol{x}_k$ is known for all units of the population U, the values

p°_{ki} are available for all $k \in U$ and we propose to use such values to define a new estimator for P_i,

$$\begin{aligned}\widehat{P}^{\mathrm{DW}}_{\mathrm{ML}i} &= N^{-1}\left(\sum_{k\in U} p^{\circ}_{ki} + \sum_{k\in s_A} d^{\circ}_k (z_{ki} - p^{\circ}_{ki}) + \sum_{k\in s_B} d^{\circ}_k (z_{ki} - p^{\circ}_{ki})\right) \\ &= N^{-1}\left(\sum_{k\in U} p^{\circ}_{ki} + \sum_{k\in s} d^{\circ}_k (z_{ki} - p^{\circ}_{ki})\right).\end{aligned} \tag{13}$$

We observe that this estimator takes the same model-assisted form as the MLGREG estimator proposed in Lehtonen and Veijanen (1998), but here it is adjusted to account for the dual frame sampling setting. The subscript *ML* stands for Multinomial-Logistic and the superscript DW stands Dual frame setting and auxiliary information available from the Whole population.

Note that we cannot compute $\sum_{k\in U} p^{\circ}_{ki}$ in (13) without knowing $\boldsymbol{x}_k$ for each $k \in U$, i.e. we need the value of the auxiliary variables for each individual in the population. This assumption can be quite restrictive; nonetheless, it can be relaxed. For example, if we have two discrete or categorical variables, we only need the population counts in the two-way contingency table. In human populations, sizes of certain demographic groups are known and are used often as auxiliary information. This is also the case in the OPIA survey and this information can be retrieved from the last column in Table 3.

An important way to incorporate available auxiliary information is given by calibration estimation (Deville and Särndal (1992)), 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 auxiliary information. We have reviewed in the previous section extension of linear calibration to the dual frame setting. Here, using the idea of model calibration introduced by Wu and Sitter (2001a), we propose the following model calibration estimator (the subscript MLC stands for Multinomial-Logistic and Calibration, and the superscript DW stands Dual frame setting and auxiliary information available from the Whole population), given by

$$\widehat{P}^{\mathrm{DW}}_{\mathrm{MLC}i} = N^{-1}\Big(\sum_{k\in s_A} w^{\circ}_k z_{ki} + \sum_{k\in s_B} w^{\circ}_k z_{ki}\Big) = N^{-1}\Big(\sum_{k\in s} w^{\circ}_k z_{ki}\Big),$$

where w°_k minimizes $\sum_{k\in s_A} G(w^{\circ}_k, d^{\circ}_k) + \sum_{k\in s_B} G(w^{\circ}_k, d^{\circ}_k) = \sum_{k\in s} G(w^{\circ}_k, d^{\circ}_k)$ for a distance measure $G(\cdot,\cdot)$ as those considered in Deville and Särndal (1992), subject to:

$$\sum_{k\in s} w^{\circ}_k p^{\circ}_{ki} = \sum_{k\in U} p^{\circ}_{ki}, \quad \sum_{k\in s_a} w^{\circ}_k = N_a, \quad \sum_{k\in s_b} w^{\circ}_k = N_b,$$

$$\sum_{k\in s_{ab}} w_k^{\circ} = \eta N_{ab} \quad \text{and} \quad \sum_{k\in s_{ba}} w_k^{\circ} = (1-\eta) N_{ab}.$$

Suppose, now, that for each unit in the population inclusion probabilities in domain *ab* are known for both frames, and not just for the frame from which the unit is selected. In this situation, the single-frame approach can also be used to propose new multinomial logistic estimators. First, we calculate $\tilde{\boldsymbol{\beta}}$ by maximizing the π-weighted log-likelihood given by

$$\ell_{\tilde{d}}(\boldsymbol{\beta}) = \sum_{i=1,\ldots,m} \sum_{k\in s} \tilde{d}_k z_{ki} \ln \mu_{ki}. \tag{14}$$

We use the new auxiliary variable $\tilde{p}_{ki} = \tilde{\mu}_{ki} = \dfrac{\exp(\boldsymbol{x}_k^{\mathrm{T}} \tilde{\boldsymbol{\beta}}_i)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_k^{\mathrm{T}} \tilde{\boldsymbol{\beta}}_r)}$ to define a new estimator (the subscript ML stands for Multinomial-Logistic and the superscript SW stands Single frame setting and auxiliary information available from the Whole population):

$$\begin{aligned} \widehat{P}_{\mathrm{ML}i}^{\mathrm{SW}} &= N^{-1}\left(\sum_{k\in U} \tilde{p}_{ki} + \sum_{k\in s_A} \tilde{d}_k (z_{ki} - \tilde{p}_{ki}) + \sum_{k\in s_B} \tilde{d}_k (z_{ki} - \tilde{p}_{ki}) \right) \\ &= N^{-1}\left(\sum_{k\in U} \tilde{p}_{ki} + \sum_{k\in s} \tilde{d}_k (z_{ki} - \tilde{p}_{ki}) \right). \end{aligned} \tag{15}$$

Note that $\tilde{d}_k$ weights are used in the formulation of the estimator (15) and also in the likelihood function (14).

Model calibration can be also used to define a single-frame estimator (the subscript MLC stands for Multinomial-Logistic and Calibration, and the superscript SW stands Single frame setting and auxiliary information available from the Whole population):

$$\widehat{P}_{\mathrm{MLC}i}^{\mathrm{SW}} = N^{-1}\Big(\sum_{k\in s_A} \tilde{w}_k z_{ki} + \sum_{k\in s_B} \tilde{w}_k z_{ki}\Big) = N^{-1}\Big(\sum_{k\in s} \tilde{w}_k z_{ki}\Big),$$

where $\tilde{w}_k$ minimizes $\sum_{k\in s_A} G(\tilde{w}_k, \tilde{d}_k) + \sum_{k\in s_B} G(\tilde{w}_k, \tilde{d}_k) = \sum_{k\in s} G(\tilde{w}_k, \tilde{d}_k)$ for a distance measure $G(\cdot,\cdot)$ satisfying the usual conditions specified in the calibration paradigm subject to:

$$\sum_{k\in s} \tilde{w}_k \tilde{p}_{ki} = \sum_{k\in U} \tilde{p}_{ki}, \quad \sum_{k\in s_a} \tilde{w}_k = N_a, \quad \sum_{k\in s_b} \tilde{w}_k = N_b \quad \text{and} \quad \sum_{k\in s_{ab}\bigcup s_{ba}} \tilde{w}_k = N_{ab}.$$

Note that when inclusion probabilities are known for both frames, it is possible to calculate single and dual frame type estimators.

4.2. Case II: Two different sets of auxiliary variables are available according the frame considered

Now we consider a different situation: the auxiliary information is available separately in each frame. In this case, for each unit $k \in \mathscr{A}$ we have an auxiliary vector $\boldsymbol{x}_{Ak}$ and for each unit $k \in \mathscr{B}$ we have another auxiliary vector $\boldsymbol{x}_{Bk}$ where the components of $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ can be different. Indeed in the OPIA survey the two sets of auxiliary variables coincide. Nonetheless, we will leave the treatment general and provide two proposals based on the dual frame approach to handle this situation as well.

In this case, we can use the available auxiliary information to fit a multinomial logistic model separately in each frame. For each $k \in \mathscr{A}$, using data from s_A we can compute

$$p_{ki}^A = \frac{\exp(\boldsymbol{x}_{Ak}^\intercal \widehat{\boldsymbol{\beta}}_i^A)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_{Ak}^\intercal \widehat{\boldsymbol{\beta}}_r^A)} \tag{16}$$

where we estimate $\boldsymbol{\beta}^A$ by maximizing $\ell_{d_A}(\boldsymbol{\beta}^A) = \sum_{i=1,\ldots,m} \sum_{k \in s_A} d_{Ak} z_{ki} \ln \mu_{ki}$. Similarly we obtain p_{ki}^B for $k \in \mathscr{B}$, and define for each $i = 1, \ldots, m$ the following regression estimator:

$$\begin{aligned}\widehat{P}_{\mathrm{ML}i}^{\mathrm{DF}} = N^{-1} \Bigg(& \sum_a p_{ki}^A + \eta \sum_{ab} p_{ki}^A + (1-\eta) \sum_{ba} p_{ki}^B + \sum_b p_{ki}^B + \\ & + \sum_{s_a} (z_{ki} - p_{ki}^A) d_{Ak} + \eta \sum_{s_{ab}} (z_{ki} - p_{ki}^A) d_{Ak} + \\ & + (1-\eta) \sum_{s_{ba}} (z_{ki} - p_{ki}^B) d_{Bk} + \sum_{s_b} (z_{ki} - p_{ki}^B) d_{Bk} \Bigg) .\end{aligned}$$

As in the previous section, the subscript ML stands for Multinomial-Logistic, while the superscript DF stands now for Dual frame setting and auxiliary information available from the Frames. To compute $\widehat{P}_{\mathrm{ML}i}^{\mathrm{DF}}$ we only need to know the total number of individuals in each domain (a, b and ab) for every possible combination of values of the auxiliary variables in the cases where discrete variables have been used as auxiliary information. In the OPIA survey this information is obtained from Table 3.

A calibration estimator in this setting can be defined under the dual frame approach as follows,

$$\widehat{P}_{\mathrm{MLC}i}^{\mathrm{DF}} = N^{-1} \Big(\sum_{k \in s_A} w_k^\star z_{ki} + \sum_{k \in s_B} w_k^\star z_{ki} \Big) = N^{-1} \Big(\sum_{k \in s} w_k^\star z_{ki} \Big), \tag{17}$$

where the subscript MLC stands for Multinomial-Logistic and Calibration, and the superscript DF stands Dual frame setting and auxiliary information available from the Frames. Weights $w_k^\star$ are such that

$$\begin{aligned}
&\min \sum_{k\in s_A} G(w_k^\star, d_{Ak}) + \sum_{k\in s_B} G(w_k^\star, d_{Bk}) \qquad \text{s.t.}\\
&\sum_{k\in s_A} w_k^\star p_{ki}^A = \sum_{k\in a} p_{ki}^A + \eta \sum_{k\in ab} p_{ki}^A,\\
&\sum_{k\in s_B} w_k^\star p_{ki}^B = (1-\eta)\sum_{k\in ba} p_{ki}^B + \sum_{k\in b} p_{ki}^B,\\
&\sum_{k\in s_a} w_k^\star = N_a, \quad \sum_{k\in s_b} w_k^\star = N_b,\\
&\sum_{k\in s_{ab}} w_k^\star = \eta N_{ab} \quad \text{and} \quad \sum_{k\in s_{ba}} w_k^\star = (1-\eta)N_{ab},
\end{aligned}$$

where p_{ki}^A are the estimated probabilities defined in (16) and p_{ki}^B are their analogous in frame B.

5. Properties of proposed estimators

To show the asymptotic properties of the proposed estimators $\hat{P}_{\text{ML}}^{\text{DW}}$, $\hat{P}_{\text{MLC}}^{\text{DW}}$, $\hat{P}_{\text{ML}}^{\text{SW}}$, $\hat{P}_{\text{MLC}}^{\text{SW}}$, $\hat{P}_{\text{ML}}^{\text{DF}}$, $\hat{P}_{\text{MLC}}^{\text{DF}}$, 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 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.

We first discuss the theoretical properties of $\hat{P}_{\text{MLC}}^{\text{DW}}$ and then move to the other estimators, because these can be dealt with using slight modifications of this more general setting. Let $\mu(\boldsymbol{x}_k, \boldsymbol{\theta}_i) = \exp(\boldsymbol{x}_k^\intercal \boldsymbol{\theta}_i)/\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_k^\intercal \boldsymbol{\theta}_r)$, for $i = 1, \ldots, m$. In order to prove our results, we make the following technical assumptions.

A1 *Let $\boldsymbol{\beta}_U$ be census level parameter estimate obtained by maximizing the likelihood $\ell_U(\boldsymbol{\beta}) = \sum_{i=1,\ldots,m} \sum_{k\in U} z_{ki} \ln \mu_{ki}$. Assume that $\boldsymbol{\beta} = \lim_{N\to\infty} \boldsymbol{\beta}_U$ exists and that $\hat{\boldsymbol{\beta}}^\circ = \boldsymbol{\beta}_U + O_p(n_N^{-1/2})$.*

A2 *For each* $\boldsymbol{x}_k$, $|\partial\mu(\boldsymbol{x}_k,\boldsymbol{\theta}_i)/\partial\boldsymbol{\theta}_i| \leq f_1(\boldsymbol{x}_k,\boldsymbol{\beta}_i)$ *for* $\boldsymbol{\theta}_i$ *in a neighborhood of* $\boldsymbol{\beta}_i$ *and* $f_1(\boldsymbol{x}_k,\boldsymbol{\beta}_i) = O(1)$, *for* $i = 1,\ldots,m$.

A3 *For each* $\boldsymbol{x}_k$, $\max_{j,j'}|\partial^2\mu(\boldsymbol{x}_k,\boldsymbol{\theta}_i)/\partial\theta_j\theta_{j'}| \leq f_2(\boldsymbol{x}_k,\boldsymbol{\beta}_i)$ *for* $\boldsymbol{\theta}_i$ *in a neighborhood of* $\boldsymbol{\beta}_i$ *and* $f_2(\boldsymbol{x}_k,\boldsymbol{\beta}_i) = O(1)$, *for* $i = 1,\ldots,m$.

A4 *The auxiliary variables* $\boldsymbol{x}$ *have bounded fourth moments.*

A5 *For any study variable* ξ *with bounded fourth moment, the sampling designs are such that for the normalized Hartley estimators of* $\bar{\xi} = N^{-1}\sum_{k\in U}\xi_k$ *a central limit theorem holds, i.e.*

$$\sqrt{n_N}(\hat{\bar{\xi}}_H - \bar{\xi}) \xrightarrow{\mathscr{L}} N(0, V(\hat{\bar{\xi}}_H)),$$

where $\hat{\bar{\xi}}_H = N^{-1}\sum_{k\in s} d_k^\circ \xi_k$ *and* $V(\hat{\bar{\xi}}_H) = V(\hat{\bar{\xi}}_a + \eta\hat{\bar{\xi}}_{ab}) + V((1-\eta)\hat{\bar{\xi}}_{ba} + \hat{\bar{\xi}}_b)$. *The latter can be consistently estimated by* $v(\hat{\bar{\xi}}_H) = v(\hat{\bar{\xi}}_a + \eta\hat{\bar{\xi}}_{ab}) + v((1-\eta)\hat{\bar{\xi}}_{ba} + \hat{\bar{\xi}}_b)$.

Assumption A1 requires consistency of parameter estimates defined by weighted estimating equations to their census level counterpart. See e.g. Binder (1983). We will first state the properties of $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$ for the Euclidean distance. In fact, in this case an analytic solution to the constrained distance minimization problem exists and is given by

$$\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i} = N^{-1}\left\{\sum_{k\in s} d_k^\circ z_{ki} + \Big(\sum_{k\in U}\tilde{p}_{ki}^\circ - \sum_{k\in s} d_k^\circ \tilde{p}_{ki}^\circ\Big)^{\mathsf{T}}\hat{\boldsymbol{\alpha}}_i^\circ\right\},$$

where $\tilde{p}_{ki}^\circ = (\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b), p_{ki}^\circ)^{\mathsf{T}}$ is a vector that contains p_{ki}° defined in (12) and a set of indicator variables – $\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b)$ – implicitly used in the benchmark constraints. In particular, $\delta_k(a)$ takes value 1 if unit $k \in U$ belongs to domain a and 0 otherwise. Then $\sum_{k\in U}\delta_k(a) = N_a$. The other indicator variables are defined similarly. In addition, $\hat{\boldsymbol{\alpha}}_i^\circ = (\sum_{k\in s} d_k^\circ \tilde{p}_{ki}^\circ \tilde{p}_{ki}^{\circ T})^{-1}(\sum_{k\in s} d_k^\circ \tilde{p}_{ki}^\circ z_{ki})$, i.e. it is the vector of coefficients of the generalized regression of z_{ki} on $\tilde{p}_{ki}^\circ$ similar to the case of classical model calibration for one frame only (see Wu and Sitter (2001a)). Then from calibration theory (see Deville and Särndal (1992)), it is well known that all other calibration estimators that use different distance functions are equivalent to $\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i}$, under additional regularity conditions on the shape of the distance function itself.

Theorem 1 *Under assumptions A1–A5,* $\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i}$ *is design* $\sqrt{n_N}$*-consistent for* P_i *in the sense that*

$$\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i} - P_i = O_p(n_N^{-1/2}),$$

and has the following asymptotic distribution

$$\frac{\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i} - P_i}{\sqrt{V_\infty(\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i})}} \xrightarrow{\mathscr{L}} N(0,1)$$

where $V_\infty(\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i}) = N^{-2}V(\hat{t}_{eiH})$ *and* $\hat{t}_{eiH} = \sum_{k\in s} d_k^\circ e_{ki}$ *is the Hartley estimator of the population total of the census-level residuals* $e_{ki} = z_{ki} - \tilde{\boldsymbol{\mu}}_{ki}^{\circ T}\boldsymbol{\alpha}_i^\circ$, *and* $\boldsymbol{\alpha}_i^\circ = (\sum_{k\in U}\tilde{\boldsymbol{\mu}}_{ki}^\circ\tilde{\boldsymbol{\mu}}_{ki}^{\circ T})^{-1}(\sum_{k\in U}\tilde{\boldsymbol{\mu}}_{ki}^\circ z_{ki})$, *where* $\tilde{\boldsymbol{\mu}}_{ki}^\circ$ *is like* $\tilde{p}_{ki}^\circ$ *but with* p_{ki}° *replaced by its population counterpart*

$$\mu_{ki}^\circ = \frac{\exp(\boldsymbol{x}_k^\intercal\boldsymbol{\beta}_{Ui})}{\sum_{r=1,\ldots,m}\exp(\boldsymbol{x}_k^\intercal\boldsymbol{\beta}_{Ur})}. \tag{18}$$

In addition, let $\hat{e}_{ki} = z_{ki} - \tilde{p}_{ki}^{\circ T}\hat{\boldsymbol{\alpha}}_i^\circ$. *Then,* $V(\hat{t}_{eiH})$ *can be consistently estimated by*

$$\begin{aligned} v(\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i}) &= N^{-2}v(\hat{t}_{\hat{e}iH}) \\ &= N^{-2}\Big\{v\Big(\textstyle\sum_{k\in s_a} d_{Ak}\hat{e}_{ki} + \eta\sum_{k\in s_{ab}} d_{Ak}\hat{e}_{ki}\Big) + \\ &\quad + v\Big((1-\eta)\textstyle\sum_{k\in s_{ba}} d_{Bk}\hat{e}_{ki} + \sum_{k\in s_b} d_{Bk}\hat{e}_{ki}\Big)\Big\}. \end{aligned} \tag{19}$$

Proof. Using the same approach developed in Montanari and Ranalli (2005) and similarly to Wu and Sitter (2001b), it is easy to show that by assumptions A1–A2 and A4–A5,

$$N^{-1}(\sum_{k\in s} d_k^\circ p_{ki}^\circ - \sum_{k\in U} p_{ki}^\circ) = O_p(n_N^{-1/2}),$$

using a first order Taylor expansion of $\mu(\boldsymbol{x}_k, \hat{\boldsymbol{\beta}}_i^\circ)$ at $\hat{\boldsymbol{\beta}}_i^\circ = \boldsymbol{\beta}_{Ui}$, and that $\hat{\boldsymbol{\alpha}}_i^\circ - \boldsymbol{\alpha}_i^\circ = O_p(n_N^{-1/2})$ because $\hat{\boldsymbol{\alpha}}_i^\circ$ is just a function of population means of variables with finite fourth moments, that can be consistently estimated by their Hartley counterparts. Using A1–A5 and a second order Taylor expansion of $\mu(\boldsymbol{x}_k, \hat{\boldsymbol{\beta}}_i^\circ)$ at $\hat{\boldsymbol{\beta}}_i^\circ = \boldsymbol{\beta}_{Ui}$,

$$N^{-1}(\sum_{k\in s} d_k^\circ p_{ki}^\circ - \sum_{k\in U} p_{ki}^\circ) = N^{-1}(\sum_{k\in s} d_k^\circ \mu_{ki}^\circ - \sum_{k\in U} \mu_{ki}^\circ) + O_p(n_N^{-1}).$$

Then,

$$\widehat{P}^{\mathrm{GDW}}_{\mathrm{MLC}i} = N^{-1}\sum_{k\in s} d_k^\circ z_{ki} + N^{-1}\Big(\sum_{k\in U}\tilde{\boldsymbol{\mu}}_{ki}^\circ - \sum_{k\in s} d_k^\circ\tilde{\boldsymbol{\mu}}_{ki}^\circ\Big)^\intercal\boldsymbol{\alpha}_i^\circ + O_p(n_N^{-1})$$

and the first part of the result is proven.

Now, from assumption A5, $v(\hat{t}_{eiH}) = V(\hat{t}_{eiH}) + o_p(n_N^{-1})$. Since $p_{ki}^{\circ} = \mu_{ki}^{\circ} + O_p(n_N^{-1/2})$, $\hat{e}_{ki} = e_{ki} + O_p(n_N^{-1/2})$ and $v(\hat{t}_{\hat{e}iH}) = v(\hat{t}_{eiH}) + o_p(n_N^{-3/2})$, then the argument follows. ■

Note that, given the asymptotic equivalence of calibration and generalized regression estimation, analytic variance estimator in (19) can be used to estimate the variance of $\hat{P}_{\mathrm{MLC}}^{\mathrm{DW}}$ also when using different distance functions.

Now, $\hat{P}_{\mathrm{ML}}^{\mathrm{DW}}$ can be seen as a particular case of $\widehat{P}_{\mathrm{MLC}i}^{GDF}$ in which $\tilde{p}_{ki}^{\circ}$ includes only p_{ki}°, and $\hat{\boldsymbol{\alpha}}_i^{\circ}$ is only a scalar and is set exactly equal to 1. Therefore, $\hat{P}_{\mathrm{ML}}^{\mathrm{DW}}$ is consistent for P_i and asymptotically normal with $V_{\infty}(\hat{P}_{\mathrm{ML}}^{\mathrm{DW}}) = N^{-2}V(\hat{t}_{eiH})$, where census-level residuals are given here by $e_{ki} = z_{ki} - \mu_{ki}^{\circ}$. Variance estimation can again be conducted by plugging sample level estimated residuals in (19) given in this case by $\hat{e}_{ki} = z_{ki} - p_{ki}^{\circ}$.

Estimator $\hat{P}_{\mathrm{MLC}}^{\mathrm{DF}}$ is in all similar to $\hat{P}_{\mathrm{MLC}}^{\mathrm{DW}}$, the only difference is in the fact that coefficient estimates for the multinomial model are obtained separately from the two frames and, therefore, we have two separate model calibration constraints. In this case the vector of auxiliary variables used in the calibration procedure can be written as $\tilde{p}_{ki}^{A,B}$ and contains p_{ki}^{A}, p_{ki}^{B} and the other indicator variables used in the benchmark constraints: for example $\tilde{p}_{ki}^{A,B} = (\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b), [\delta_k(a) + \delta_k(ab)]p_{ki}^{A}, [\delta_k(b) + \delta_k(ba)]p_{ki}^{B})^{\intercal}$.

To encompass this situation, it is enough to change assumption A1 accordingly and assume that the two sets of population parameters $\boldsymbol{\beta}^A$ and $\boldsymbol{\beta}^B$ are consistently estimated by $\hat{\boldsymbol{\beta}}^A$ and $\hat{\boldsymbol{\beta}}^B$ and that these samples fits and the finite population fits share a common finite limit. Then, it is easy to show that $\hat{P}_{\mathrm{MLC}}^{\mathrm{DF}}$ is design consistent and the variance of its asymptotic normal distribution can again be written in terms of the variance of the population total of residuals. In particular, $V_{\infty}(\widehat{P}_{\mathrm{MLC}i}^{GDF}) = N^{-2}V(\hat{t}_{eiH})$ and $\hat{t}_{eiH} = \sum_{k\in s} d_k^{\circ} e_{ki}$ is the Hartley estimator of the population total of the census-level residuals given here by $e_{ki} = z_{ki} - (\tilde{\boldsymbol{\mu}}^{A,B})_{ki}^{\intercal}\boldsymbol{\alpha}_i$, where $\tilde{\boldsymbol{\mu}}_{ki}^{A,B}$ is like $\tilde{p}_{ki}^{A,B}$ but with p_{ki}^{A} and p_{ki}^{B} replaced by their population counterparts, similarly to (18). Analytic variance estimation can be conducted by using sample level estimates of the residuals. In particular, by using $\hat{e}_{ki} = z_{ki} - (\tilde{p}_{ki}^{A,B})^{\intercal}\hat{\boldsymbol{\alpha}}_i$ in formula (19).

Now, similarly as for $\hat{P}_{\mathrm{ML}}^{\mathrm{DW}}$ and $\hat{P}_{\mathrm{MLC}}^{\mathrm{DW}}$, $\hat{P}_{\mathrm{ML}}^{\mathrm{DF}}$ can be seen as a particular case of $\widehat{P}_{\mathrm{MLC}i}^{GDF}$ in which $\tilde{p}_{ki}^{\circ}$ includes only $p_{ki}^{A,B}$, with $p_{ki}^{A,B} = p_{ki}^{A}$ if $k \in s_A$ and $p_{ki}^{A,B} = p_{ki}^{B}$ if $k \in s_B$, and $\hat{\boldsymbol{\alpha}}_i^{\circ}$ is again a scalar here and its value is set exactly equal to 1. Therefore, it is consistent for P_i and asymptotically normal with $V_{\infty}(\hat{P}_{\mathrm{ML}}^{\mathrm{DF}}) = N^{-2}V(\hat{t}_{eiH})$, where census-level residuals are given here by $e_{ki} = z_{ki} - \mu_{ki}^{A,B}$, and $\mu_{ki}^{A,B}$ is the census level fit corresponding to $p_{ki}^{A,B}$. Variance estimation can again be conducted by using sample level estimated residuals in equation (19) given by $\hat{e}_{ki} = z_{ki} - p_{ki}^{A}$ if $k \in s_A$ and $\hat{e}_{ki} = z_{ki} - p_{ki}^{B}$ if $k \in s_B$.

The calibration estimator $\hat{P}_{\mathrm{MLC}}^{\mathrm{SW}}$ is very similar to $\hat{P}_{\mathrm{MLC}}^{\mathrm{DW}}$, the only differences are (i) in the set of basic design weights employed in the calibration procedure: for $\hat{P}_{\mathrm{MLC}}^{\mathrm{SW}}$ we use $\tilde{d}_k$, and (ii) p_{ki}° is replaced by $\tilde{p}_{ki}$ in the definition of the vector $\tilde{p}_{ki}^{\circ}$. Once these changes are incorporated across assumption A1, and assumption A5 reflects the fact that we are now dealing with Bankier-Kalton-Anderson type estimators, instead of Hartley estimators, then all the results can be proven. The variance of the asymptotic distribution

of $\hat{P}^{\mathrm{SW}}_{\mathrm{MLC}}$ is given by $V_\infty(\widehat{P}^{\mathrm{GSW}}_{\mathrm{MLC}i}) = N^{-2}V(\hat{t}_{ei})$ and $\hat{t}_{ei} = \sum_{k\in s}\tilde{d}_k e_{ki}$ is the single-frame estimator of the population total of the census-level residuals $e_{ki} = z_{ki} - \tilde{\boldsymbol{\mu}}_{ki}^{\mathrm{T}}\boldsymbol{\alpha}_i$, and where $\tilde{\boldsymbol{\mu}}_{ki}$ is like $\tilde{p}_{ki}$ but with p_{ki} replaced by its population counterpart

$$\mu_{ki} = \frac{\exp(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{\beta}_{Ui})}{\sum_{r=1,\ldots,m}\exp(\boldsymbol{x}_k^{\mathrm{T}}\boldsymbol{\beta}_{Ur})}.$$

In addition, let $\hat{e}_{ki} = z_{ki} - \tilde{p}_{ki}^{T}\hat{\boldsymbol{\alpha}}_i$. Then, $V(\hat{t}_{ei})$ can be consistently estimated so that $v(\widehat{P}^{\mathrm{GSW}}_{\mathrm{MLC}i}) = N^{-2}v(\hat{t}_{\hat{e}i})$.

6. Selection of the optimal weight

In the previous sections we have considered a fixed value $0 < \eta < 1$. Selection of parameter η is an important issue in dual frame estimators, because the efficiency of the estimator relies heavily on this value (see Lohr (2009) for a review). Hartley (1962) proposed choosing η to minimize the variance of the estimator in (2). Using the same idea, we can derive the optimal value of η for each proposed multinomial logistic estimator by minimizing its asymptotic variance with respect to η. However, as the optimal value for the Hartley estimator, such optimal values would depend on unknown population quantities, such as variances and covariances that, when estimated from sample data, would make the final estimator depend on the values of the variable of interest. This implies a need to recompute an optimal η for each value $i = 1,\ldots,m$ and for each variable of interest y, which will be inconvenient in practice for statistical agencies conducting surveys with several variables, other than introducing a lack in coherence among estimates that is particularly relevant when dealing with multinomial outcomes (namely, $\sum_i \widehat{P}_i$ can be $\neq 1$).

Skinner and Rao (1996) suggested choosing

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

or alternatively

$$\eta_{SR2} = \frac{V(\hat{N}_{ba})}{V(\hat{N}_{ba}) + V(\hat{N}_{ab})},$$

being $V(\hat{N}_{ab})$ and $V(\hat{N}_{ba})$ the variances of the estimated sizes of domain ab based on samples s_A and s_B respectively. These two proposals provide a value for η that does not depend on the sample values of y. In this way, resulting estimator uses the same η for all variables of interest, even if variances $V(\hat{N}_{ab})$ and $V(\hat{N}_{ba})$ are unknown and must be estimated from the data.

Brick et al. (2006) propose using the simple value $\eta = 1/2$ in their dual-frame study 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 (see, for example, Mecatti (2007)). Another simple choice for η is given by $\frac{N_B/n_B}{N_A/n_A+N_B/n_B}$ (see Skinner and Rao (1996) or Lohr and Rao (2000)).

7. Jackknife variance estimation

In this section we explore the possibility of using jackknife methods to estimate the variance of the proposed estimators as an alternative to the analytic variance estimators considered in Section 5. 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 (2003) for an introduction to jackknife). For the sake of brevity, in this section all estimators are denoted by $\hat{P}_i, i = 1, \cdots, m$.

If we consider a non clustered and non stratified design, the jackknife estimator for the variance of $\hat{P}_i$ may be given by

$$v_J(\hat{P}_i) = V_J^A + V_J^B = \frac{n_A - 1}{n_A} \sum_{g \in s_A} (\hat{P}_i^A(g) - \overline{P}_i^A)^2 + \frac{n_B - 1}{n_B} \sum_{j \in s_B} (\hat{P}_i^B(j) - \overline{P}_i^B)^2 \qquad (20)$$

where $\hat{P}_i^A(g)$ is the value taken by estimator $\hat{P}_i$ after dropping unit g from s_A and $\overline{P}_i^A$ is the average of $\hat{P}_i^A(g)$ values. Each value $\hat{P}_i^A(g)$ is computed by fitting a new model that does not consider the $g-th$ sample unit. $\hat{P}_i^B(j)$ and $\overline{P}_i^B$ are defined similarly.

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, stratum l has N_{Bl} observation units of which n_{Bl} are sampled. Then, a jackknife variance estimator of $\hat{P}_i$ is given by

$$v_J^{st}(\hat{P}_i) = V_J^{stA} + V_J^{stB} =$$

$$= \sum_{h=1}^{H} \frac{n_{Ah} - 1}{n_{Ah}} \sum_{g \in s_{Ah}} (\hat{P}_i^A(hg) - \overline{P}_i^{Ah})^2 + \sum_{l=1}^{L} \frac{n_{Bl} - 1}{n_{Bl}} \sum_{j \in s_{Bl}} (\hat{P}_i^B(lj) - \overline{P}_i^{Bl})^2, \qquad (21)$$

where $\hat{P}_i^A(hg)$ is the value taken by estimator $\hat{P}_i$ after dropping unit g of stratum h from sample s_{Ah}, $\overline{P}_i^{Ah}$ is the average of these n_{Ah} values; $\hat{P}_i^B(lj)$ and $\overline{P}_i^{Bl}$ 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.

Alternatively, a finite-population correction can be considered, as described in Ranalli et al. (2015), resulting in the following jackknife variance estimators:

$$v_{Jc}(\hat{P}_i) = \frac{n_A - 1}{n_A}(1-\overline{\pi}_A)\sum_{g\in s_A}(\hat{P}_i^A(g) - \overline{P}_i^A)^2 + \frac{n_B - 1}{n_B}(1-\overline{\pi}_B)\sum_{j\in s_B}(\hat{P}_i^B(j) - \overline{P}_i^B)^2 \quad (22)$$

for non stratified designs in frames, where $\overline{\pi}_A = \frac{1}{n_A}\sum_{k\in s_A}\pi_{Ak}$ and similarly for $\overline{\pi}_B$, and

$$\begin{aligned} v_{Jc}^{st}(\hat{P}_i) = &\sum_{h=1}^{H}\frac{n_{Ah}-1}{n_{Ah}}(1-\overline{\pi}_{Ah})\sum_{g\in s_{Ah}}(\hat{P}_i^A(hg) - \overline{P}_i^{Ah})^2 \\ &+\sum_{l=1}^{L}\frac{n_{Bl}-1}{n_{Bl}}(1-\overline{\pi}_{Bl})\sum_{j\in s_{Bl}}(\hat{P}_i^B(lj) - \overline{P}_i^{Bl})^2 \end{aligned} \quad (23)$$

for a stratified design in each frame, where $\overline{\pi}_{Ah} = \frac{1}{n_{Ah}}\sum_{k\in s_{Ah}}\pi_{Ak}$ and similarly for $\overline{\pi}_{Bl}$.

A non clustered sampling design is assumed subsequently. No new principles are involved in the application of jackknife methodology to clustered samples. We simply work with the ultimate cluster rather than elementary units (see e.g. Wolter (2003)).

8. Monte Carlo simulation experiments

For our simulation study we use the `hsbdemo` data set (`http://www.ats.ucla.edu/stat/data/hsbdemo.dta`). The data set contains variables on 200 students. The outcome variable is `prog`, program type, a three-level categorical variable whose categories are `academic`, `general`, `vocation`. The predictor variables are social economic status, `ses`, a three-level categorical variable and a mathematical score, `math`, a continuous variable. We estimate a multinomial logistic regression model. We create a new data set with 50 copies of the predictor variables `ses` and `math` and with the predicted values for the variable `prog` (the category with highest probability). The simulated populations, namely POP1, have, therefore, dimension $N = 10\,000$.

Units are randomly assigned to the two frames, A and B, according to three different scenarios depending on the overlap domain size N_{ab}. We first generate N normal random numbers, $\varepsilon_k, k = 1,\ldots,N$ and data is sorted by such random numbers. Then, the first N_a records of the ordered dataset are considered as the values of the domain a, the N_b subsequent records as the values belonging to domain b and the last N_{ab} records as the values of the domain ab. The first scenario has a *small* overlap domain size N_{ab}=1 000 and the resulting sizes of the two frames are N_A=6 000 and N_B=5 000. The second and the third scenario have respectively *medium* and *large* overlap domain size. The resulting frame sizes in the second scenario are given by N_A=6 000 and N_B=7 000 and the overlap domain size is N_{ab}=3 000, while for the third scenario we have N_A=8 000, N_B=7 000 and N_{ab}=5 000. In POP1, we compute all estimators using as auxiliary information `ses` and `math`.

On the other hand, POP2 is built first by assigning units to the frames and second by fitting a multinomial logistic regression model separately in each frame. In frame A, `ses`

and `math` have been considered as auxiliary variables and in frame B the auxiliary variables are `ses` and `write` (a score in writing). To be able to fit a separated model in each frame we consider that the units composing the overlap domain can be equally divided into two groups, each one coming from a frame. So half of the overlap domain units are used to fit a multinomial logistic regression model in frame A and the remaining ones are considered when fitting the multinomial logistic model in frame B. POP2 is built with the predicted values from the two multinomial logistic model. In this population, we compute $\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{SW}}_{\mathrm{MLC}}$ estimators using as x-variable `ses` (Case I), and $\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{MLC}}$ estimators using as x_A-variables `ses` and `math` and as x_B-variables `ses` and `write` (Case II).

Samples of schools from frame A are selected by means of Midzuno sampling, with inclusion probabilities proportional to the size of the school the student belongs to. All students in the selected schools are included in the sample. The variable `cid` is an indicator of school. Samples from frame B are selected by means of simple random sampling. For each scenario, we draw a combination of sample sizes for frame A and frame B, as follows: $n_A = 180$ and $n_B = 232$.

We have two populations, three sizes of the overlap domain and different sets of auxiliary variables.

We compute the BKA estimator in (9), for the purpose of comparison. The Pseudo Empirical Likelihood estimator (PEL) proposed in Rao and Wu (2010) and the dual frame and the single frame calibration estimator ($\hat{P}_{\mathrm{CalDF}}$ and $\hat{P}_{\mathrm{CalSF}}$) proposed in Ranalli et al. (2015) are also computed using the auxiliary information as previously mentioned (in POP1 `ses` and `math` for both estimators and in POP2 as x_A-variable `ses` and `math` and as x_B-variable `ses` and `write` for $\hat{P}_{\mathrm{CalDF}}$ estimator and as x-variable `ses` for $\hat{P}_{\mathrm{CalSF}}$ estimator). When needed (and for comparative purposes) the value of η has been estimated using $\eta = v(\hat{N}_{ba})/(v(\hat{N}_{ab}) + v(\hat{N}_{ba}))$ (see for example Rao and Wu (2010)) for all compared estimators, where $v(\hat{N}_{ab})$ is an estimate of the variance of the Horvitz-Thompson estimator $\hat{N}_{ab}$ for the size of overlap domain, and similarly for $v(\hat{N}_{ba})$.

For each estimator, we compute the percent relative bias $\mathrm{RB}\% = 100 * E_{\mathrm{MC}}(\hat{Y} - Y)/Y$, the percent relative mean squared error $\mathrm{RMSE}\% = 100 * E_{\mathrm{MC}}[(\hat{Y} - Y)^2]/Y^2$, based on 1000 simulation runs, for each category of the main variable `prog`.

The percent relative biases are negligible in all cases (the results on RB are not included for brevity), so efficiency comparisons can be based on variances. Table 4 displays the relative efficiency of proposed estimators with respect to BKA estimator. From this table we can see that, consistently with theoretical findings, the performance in terms of efficiency of the estimators is essentially driven by the model employed. When the auxiliary varibles are used in a calibration process using a linear model ($\hat{P}_{CalSF}$, $\hat{P}_{CalDF}$) or through a pseudo-empirical likelihood method (PEL), the efficiency increases with respect to the BKA estimator, which does not use auxiliary information or any model. As expected, a most effective situation arises when the auxiliary variables are also used through a multinomial model ($\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{MLC}}$).

Table 4: *Relative efficiency (respect to the BKA estimator) of compared estimators. POP1 and POP2.*

	POP1			POP2		
	acad.	gen.	voc.	acad.	gen.	voc.
			Medium			
$\hat{P}_{\text{BKA}}$	100.00	100.00	100.00	100.00	100.00	100.00
$\hat{P}_{\text{CalSF}}$	149.94	142.21	132.30	152.77	145.10	129.26
$\hat{P}_{\text{PEL}}$	217.89	135.87	177.26	175.94	146.75	148.75
$\hat{P}_{\text{CalDF}}$	213.91	134.83	175.14	175.03	146.84	147.59
$\hat{P}^{\text{DW}}_{\text{ML}}$	347.02	181.43	252.42	204.46	194.97	148.32
$\hat{P}^{\text{DW}}_{\text{MLC}}$	356.87	181.05	258.60	209.29	192.64	153.29
$\hat{P}^{\text{SW}}_{\text{ML}}$	348.12	181.25	252.44	205.63	194.71	148.82
$\hat{P}^{\text{SW}}_{\text{MLC}}$	358.10	180.97	258.85	210.22	192.32	153.70
$\hat{P}^{\text{DF}}_{\text{ML}}$	350.18	187.65	257.22	207.83	251.93	147.44
$\hat{P}^{\text{DF}}_{\text{MLC}}$	358.93	186.31	263.52	214.76	250.13	153.44
			Small			
$\hat{P}_{\text{BKA}}$	100.00	100.00	100.00	100.00	100.00	100.00
$\hat{P}_{\text{CalSF}}$	155.30	137.56	140.60	152.77	142.46	137.70
$\hat{P}_{\text{PEL}}$	232.55	147.36	198.25	179.24	149.26	158.30
$\hat{P}_{\text{CalDF}}$	210.50	134.54	179.08	182.73	150.09	160.65
$\hat{P}^{\text{DW}}_{\text{ML}}$	331.43	163.16	247.64	165.45	146.32	157.70
$\hat{P}^{\text{DW}}_{\text{MLC}}$	353.76	163.06	265.66	176.59	146.83	166.11
$\hat{P}^{\text{SW}}_{\text{ML}}$	331.75	163.33	248.08	166.09	146.83	157.60
$\hat{P}^{\text{SW}}_{\text{MLC}}$	353.77	163.17	265.85	176.78	146.99	165.93
$\hat{P}^{\text{DF}}_{\text{ML}}$	343.94	164.70	257.75	170.24	150.15	154.31
$\hat{P}^{\text{DF}}_{\text{MLC}}$	365.15	163.94	275.28	184.50	150.24	164.51
			Large			
$\hat{P}_{\text{BKA}}$	100.00	100.00	100.00	100.00	100.00	100.00
$\hat{P}_{\text{CalSF}}$	147.60	130.53	138.13	152.25	121.61	125.29
$\hat{P}_{\text{PEL}}$	193.48	124.99	173.21	163.71	142.12	149.74
$\hat{P}_{\text{CalDF}}$	192.10	125.72	170.56	165.55	153.62	161.09
$\hat{P}^{\text{DW}}_{\text{ML}}$	354.00	161.79	256.45	303.59	118.57	269.38
$\hat{P}^{\text{DW}}_{\text{MLC}}$	371.74	161.23	266.64	307.98	123.76	282.16
$\hat{P}^{\text{SW}}_{\text{ML}}$	356.73	161.87	257.40	302.59	119.33	269.14
$\hat{P}^{\text{SW}}_{\text{MLC}}$	375.21	161.38	267.54	306.81	124.75	281.93
$\hat{P}^{\text{DF}}_{\text{ML}}$	362.07	168.39	265.88	344.86	130.46	370.90
$\hat{P}^{\text{DF}}_{\text{MLC}}$	376.11	167.22	274.78	348.03	137.80	379.38

Table 5: *Length reduction (in percent, %) of proposed estimator with respect to linear calibration estimators using the same amount of auxiliary information ($\hat{P}^{DW}_{ML}$, $\hat{P}^{DW}_{MLC}$, $\hat{P}^{SW}_{ML}$ and $\hat{P}^{SW}_{MLC}$ have been compared with $\hat{P}_{CalSF}$ and $\hat{P}^{DF}_{ML}$ and $\hat{P}^{DF}_{MLC}$ have been compared with $\hat{P}_{CalDF}$). Coverage (in percent, %) of jackknife confidence intervals. POP1.*

	Length reduction			Cov		
	acad.	gen.	voc.	acad.	gen.	voc.
			Medium			
$\hat{P}^{DW}_{ML}$	10.31	25.44	30.91	94.5	93.9	94.9
$\hat{P}^{DW}_{MLC}$	9.90	28.28	32.78	95.2	93.9	94.5
$\hat{P}^{SW}_{ML}$	10.59	25.73	31.18	94.8	94.1	95.0
$\hat{P}^{SW}_{MLC}$	9.95	28.34	32.82	95.0	93.8	94.5
$\hat{P}^{DF}_{ML}$	8.83	33.04	16.41	95.8	96.0	95.5
$\hat{P}^{DF}_{MLC}$	8.11	35.23	18.24	95.9	95.3	95.1
			Small			
$\hat{P}^{DW}_{ML}$	9.14	23.76	28.25	95.0	93.2	95.2
$\hat{P}^{DW}_{MLC}$	8.78	26.86	30.41	94.1	93.4	93.6
$\hat{P}^{SW}_{ML}$	9.43	24.04	28.52	94.5	93.5	94.0
$\hat{P}^{SW}_{MLC}$	8.81	26.89	30.43	94.8	92.5	94.2
$\hat{P}^{DF}_{ML}$	6.98	24.64	13.09	96.3	95.0	95.9
$\hat{P}^{DF}_{MLC}$	6.30	27.15	15.32	96.6	94.6	95.1
			Large			
$\hat{P}^{DW}_{ML}$	10.11	25.45	30.71	94.2	93.5	93.9
$\hat{P}^{DW}_{MLC}$	9.34	28.24	32.38	94.1	93.4	93.6
$\hat{P}^{SW}_{ML}$	10.64	25.94	31.14	94.5	93.5	94.0
$\hat{P}^{SW}_{MLC}$	9.71	28.51	32.62	94.8	92.5	94.2
$\hat{P}^{DF}_{ML}$	10.18	35.37	17.96	96.3	95.0	95.9
$\hat{P}^{DF}_{MLC}$	9.29	37.39	19.45	96.6	94.6	95.1

In general, the best results in efficiency are achieved by the $\hat{P}^{DF}_{MLC}$ estimator and the efficiency increases as the size of the overlap domain increases, particularly for POP2. As a consequence of the ignorability of the frames the units belong to when modelling the relation between the response and the auxiliary variables, there is not a relevant difference in efficiency between estimators using a multinomial model in the whole population and estimators using a multinomial model in each frame.

We now turn to the evaluation of the precision of the proposed estimators by means of confidence intervals. We obtain the 95% confidence intervals based on a normal distribution and the jackknife variance estimator proposed in Section 7 with finite-population correction. Table 5 shows the average length reduction of 95% confidence intervals and

Table 6: *Relative efficiency (respect to the BKA estimator) of compared estimator for* $\hat{\eta}_{SR2} = v(\hat{N}_{ba})/(v(\hat{N}_{ab}) + v(\hat{N}_{ba}))$, $\hat{\eta}_{SR} = N_a N_B v(\hat{N}_{ba})/(N_b N_A v(\hat{N}_{ab}) + N_a N_B v(\hat{N}_{ba}))$ *and* $\eta_{1/2} = \frac{1}{2}$. *Overlap domain size* Medium.

		POP1			POP2		
		acad.	gen.	voc.	acad.	gen.	voc.
$\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$	$\hat{\eta}_{SR2}$	347.02	181.43	252.42	204.46	194.97	148.32
	$\hat{\eta}_{SR}$	348.45	181.32	252.88	205.14	194.69	148.71
	$\eta_{1/2}$	347.27	181.30	252.57	204.69	194.91	148.32
$\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$	$\hat{\eta}_{SR2}$	356.87	181.05	258.60	209.29	192.64	153.29
	$\hat{\eta}_{SR}$	358.65	181.01	259.21	209.78	192.36	153.62
	$\eta_{1/2}$	357.11	180.91	258.76	209.48	192.54	153.26
$\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$	$\hat{\eta}_{SR2}$	350.18	187.65	257.22	207.83	251.93	147.44
	$\hat{\eta}_{SR}$	351.57	187.70	257.90	207.85	249.31	147.45
	$\eta_{1/2}$	350.34	187.45	257.33	208.03	251.91	147.50
$\hat{P}^{\mathrm{DF}}_{\mathrm{MLC}}$	$\hat{\eta}_{SR2}$	358.93	186.31	263.52	214.76	250.13	153.44
	$\hat{\eta}_{SR}$	360.76	186.46	264.35	214.57	247.50	153.26
	$\eta_{1/2}$	215.02	250.07	153.52	182.44	148.19	163.36

the empirical coverage probability over 1000 simulation runs in each category of the main variable. The confidence interval lengths of proposed estimators have been compared with the confidence interval lengths of their linear calibration counterparts using the same amount of auxiliary information. That is, $\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{SW}}_{\mathrm{MLC}}$ have been compared with $\hat{P}_{\mathrm{CalSF}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{MLC}}$ have been compared with $\hat{P}_{\mathrm{CalDF}}$.

From Table 5 we conclude that all the proposed estimators considerably reduce the length of the confidence intervals obtained, with respect to the linear calibration estimators. The empirical coverage is very close to the nominal level. It is observed that the estimates based on the joint estimation of the parameter $\boldsymbol{\beta}$ ($\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{DW}}_{\mathrm{MLC}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{SW}}_{\mathrm{MLC}}$) have a somewhat lower coverage than the others.

Looking at the effect of the choice of η (in relative bias and relative mean squared error), we have repeated the simulation study (for all populations and scenarios) using alternative values for η. In particular, other than that used previously, i.e.

$$\eta_{SR2} = \frac{v(\hat{N}_{ba})}{v(\hat{N}_{ba}) + v(\hat{N}_{ab})},$$

we have considered a fixed value $\eta = \frac{1}{2}$ and one estimated following Skinner and Rao (1996)

$$\eta_{SR} = \frac{N_a N_B v(\hat{N}_{ba})}{N_a N_B v(\hat{N}_{ba}) + N_b N_A v(\hat{N}_{ab})}.$$

See Section 6. for details and guidelines on choosing a value for η. Table 6 shows (only when the overlap domain size is *Medium*, for space reason) that there is a little effect of these three different estimates for η on the behaviour of the considered estimators. We can conclude that the available auxiliary information and the way in which it is included in the estimation procedure play a much more relevant role than the choice of a value for η.

9. Application to the survey on opinions and attitudes of the Andalusian population regarding immigration (OPIA) 2013

To examine the performance of the proposed estimation methods in practice, we have applied them to the dataset from the OPIA survey. The main variable in this study is related to the "attitude towards immigration". The variable is the answer to the following question: *And in relation to the number of immigrants currently living in Andalusia, do you think there are ...?: Too many, A reasonable number, Too few, No reply*.

We have considered the same set of auxiliary variables (`sex` and `age`) in the two frames. To incorporate information about sex into estimation process two indicator variables (one for males and another one for females) were created. Similarly, four age classes were established and each respondent was assigned to one of them. Corresponding indicator variables were used, then, for the analysis. Necessary population information about these variables for calculating proposed estimators is displayed in Table 3. Note that both auxiliary variables `sex` and `age` are available from the two frames. In this case, the population counts in the two-way contingence table are known in each domain.

Table 7 shows point and jackknife confidence estimation for proposed estimators. Length reduction in jackknife confidence interval for each estimator regarding same interval for BKA estimator is also displayed. In keeping with results obtained from simulation experiments, reduction is quite significative for all estimators whatever the category of the main variable. The calibration approach achieves most important reductions in length, with single frame calibration presenting the best results. On the other hand, using $\hat{P}^{\mathrm{DW}}_{\mathrm{ML}}$, $\hat{P}^{\mathrm{SW}}_{\mathrm{ML}}$ and $\hat{P}^{\mathrm{DF}}_{\mathrm{ML}}$ estimators the length reduction is less noticeable.

Table 8 shows point estimation for proposed estimators by sex and age. Analyzing results by gender, it is noticeable that there are more males than females thinking that there are too many immigrants in Andalusia and that females are more reticent to answer the question than males.

On the other hand, it is worth noting that perception that there are too many immigrants in Andalusia increases together with age. So, while most of the people in the 18-29 age group think that the number of immigrants in Andalusia is reasonable, most part of people aged 45 years or over think that there are too many. The age group where the non-response is higher is the one including people aged 60 years or over.

Table 7: *Point and 95% confidence level estimation of proportions using several methods for Jackknife variance estimation. Length reduction (in percent, %) respect to the BKA estimator. Main variable: "Amount of immigration".*

In relation to the number of immigrants currently living in Andalusia, do you think there are ...?					
Estimator	PROP	LB	UB	LEN	Length reduction
			Too many		
$\hat{P}^{DW}_{ML}$	42.75	39.76	45.74	5.98	14.33
$\hat{P}^{DW}_{MLC}$	41.23	38.78	43.68	4.90	29.80
$\hat{P}^{SW}_{ML}$	42.89	39.94	45.84	5.90	15.47
$\hat{P}^{SW}_{MLC}$	41.41	39.03	43.79	4.76	31.81
$\hat{P}^{DF}_{ML}$	42.61	39.64	45.58	5.94	14.90
$\hat{P}^{DF}_{MLC}$	41.16	38.67	43.65	4.98	28.65
			A reasonable number		
$\hat{P}^{DW}_{ML}$	45.24	42.27	48.20	5.93	12.28
$\hat{P}^{DW}_{MLC}$	46.57	44.11	49.03	4.92	27.22
$\hat{P}^{SW}_{ML}$	45.09	42.17	48.01	5.84	13.61
$\hat{P}^{SW}_{MLC}$	46.40	44.02	48.78	4.76	29.59
$\hat{P}^{DF}_{ML}$	45.45	42.49	48.41	5.92	12.43
$\hat{P}^{DF}_{MLC}$	46.68	44.17	49.18	5.01	25.89
			Too few		
$\hat{P}^{DW}_{ML}$	6.06	4.55	7.58	3.03	15.36
$\hat{P}^{DW}_{MLC}$	5.77	4.58	6.97	2.39	33.24
$\hat{P}^{SW}_{ML}$	6.05	4.56	7.54	2.98	16.76
$\hat{P}^{SW}_{MLC}$	5.76	4.61	6.91	2.30	35.75
$\hat{P}^{DF}_{ML}$	6.13	4.62	7.64	3.02	15.64
$\hat{P}^{DF}_{MLC}$	5.63	4.46	6.80	2.34	34.64
			No reply		
$\hat{P}^{DW}_{ML}$	5.95	4.65	7.25	2.60	12.75
$\hat{P}^{DW}_{MLC}$	6.43	5.27	7.58	2.31	22.48
$\hat{P}^{SW}_{ML}$	5.96	4.67	7.25	2.58	13.42
$\hat{P}^{SW}_{MLC}$	6.43	5.30	7.56	2.26	24.16
$\hat{P}^{DF}_{ML}$	5.80	4.51	7.10	2.59	13.09
$\hat{P}^{DF}_{MLC}$	6.54	5.33	7.74	2.41	19.13

Table 8: *Point estimation of proportions by sex and age. Main variable: "Amount of immigration".*

	In relation to the number of immigrants currently living in Andalusia, do you think there are ...?						
Estimator	ALL	MALES	FEMALES	18-29	30-44	45-59	≥ 60
			Too many				
$\hat{P}^{DW}_{ML}$	42.75	46.46	39.15	32.46	44.29	46.03	45.14
$\hat{P}^{DW}_{MLC}$	41.23	43.64	38.97	30.97	42.07	43.31	46.58
$\hat{P}^{SW}_{ML}$	42.89	46.74	39.11	32.76	43.89	46.44	45.85
$\hat{P}^{SW}_{MLC}$	41.41	43.79	39.19	31.55	41.61	43.87	45.77
$\hat{P}^{DF}_{ML}$	42.61	44.45	39.16	31.99	41.69	43.56	48.13
$\hat{P}^{DF}_{MLC}$	41.16	43.55	38.96	30.01	42.14	43.28	48.56
			A reasonable number				
$\hat{P}^{DW}_{ML}$	45.24	42.31	48.10	59.82	40.71	40.72	44.47
$\hat{P}^{DW}_{MLC}$	46.57	44.39	48.74	61.97	44.44	42.72	43.25
$\hat{P}^{SW}_{ML}$	45.09	42.04	48.11	59.62	40.90	40.68	43.70
$\hat{P}^{SW}_{MLC}$	46.40	44.14	48.63	61.49	44.67	42.64	43.61
$\hat{P}^{DF}_{ML}$	45.45	44.02	48.35	60.42	43.98	42.81	42.11
$\hat{P}^{DF}_{MLC}$	46.68	44.59	48.78	63.21	44.46	42.56	41.65
			Too few				
$\hat{P}^{DW}_{ML}$	6.06	6.75	5.35	3.77	9.84	6.18	2.82
$\hat{P}^{DW}_{MLC}$	5.77	6.68	4.92	3.29	7.58	6.73	2.80
$\hat{P}^{SW}_{ML}$	6.05	6.64	5.47	3.79	9.89	6.12	2.83
$\hat{P}^{SW}_{MLC}$	5.76	6.67	4.92	3.39	7.62	6.66	2.95
$\hat{P}^{DF}_{ML}$	6.13	6.58	5.11	3.50	8.17	6.37	2.39
$\hat{P}^{DF}_{MLC}$	5.63	6.46	4.81	2.92	7.46	6.77	2.35
			No reply				
$\hat{P}^{DW}_{ML}$	5.95	4.47	7.39	3.95	5.16	7.06	7.56
$\hat{P}^{DW}_{MLC}$	6.43	5.28	7.37	3.76	5.91	7.24	7.37
$\hat{P}^{SW}_{ML}$	5.96	4.58	7.31	3.83	5.32	6.76	7.62
$\hat{P}^{SW}_{MLC}$	6.43	5.41	7.26	3.57	6.10	6.84	7.67
$\hat{P}^{DF}_{ML}$	5.80	4.95	7.38	4.09	6.15	7.25	7.36
$\hat{P}^{DF}_{MLC}$	6.54	5.39	7.45	3.86	5.93	7.39	7.44

10. Conclusions

Data collected from surveys are often organized into discrete categories. Analyzing such categorical data from a complex survey often requires specialized techniques. To improve the accuracy of estimation procedures, a survey statistician often makes use of the auxiliary data available from administrative registers and other sources.

Generalized regression is a popular design-based method used in the production of descriptive statistics from survey data. Although the generalized regression estimator is design-consistent regardless of the form of the assisting model, a linear model is not the best choice for multinomial response variables. For such variables we introduce a class of multinomial logistic generalized regression estimators when data are obtained from samples from different frames.

We introduce a new approach to the model-assisted estimation of population class of frequencies in dual frame surveys. We propose a class of logistic estimators based on multinomial logistic models describing the joint distribution of the category indicators in the total population or in each frame separately. We also consider different ways of combining estimates coming from the two frames.

The type of sample design used in practice drives the user to choose between Dual-Frame or Single-Frame approaches. The Single-Frame approach requires additional information in the overlapping domain that is not always easy to take in practical applications.

As for calibration, it seems clear that the better for efficiency is to incorporate it, regardless of whether or not a logistics model is used. As for the model, apart from the advantage provided by the fact that the estimates of proportions for each category add to one, our simulation study suggests that it is preferable to use it. As for the type of model, in most practical applications it will be almost entirely forced, depending on the auxiliary information available and, more specifically, on the availability of auxiliary variable totals for domains, for frames or for the entire population.

To compute the proposed estimators, we have assumed to know the values of auxiliary variables for each individual in the population, which can be quite a restrictive assumption. Indeed, to compute the proposed estimators we need to know the count of each value of the auxiliary variable vector in the population. This is a very frequent situation that arises, for example, when categorical variables (as the gender or the professional status of the individual) or quantitative categorized variables (as the age of the individual, grouped in classes) are used as auxiliary information in a survey. In this context, we do not have a complete list of individuals but still the proposed estimators can be computed since the population information needed can be found in databases of national statistical organisms. In fact, in this case, we only need to know the population count in the multi-way contingency table. This is also the situation in the application to data from the survey on opinions and attitudes of the Andalusian population regarding immigration explored in Section 9.

Here we have considered two frames. 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) and 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.

Acknowledgements

This study was partially supported by Ministerio de Economía y Competitividad (grant MTM2012-35650 and FPU grant program, Spain), by Consejería de Economía, Innovación, Ciencia y Empleo (grant SEJ2954, Junta de Andalucía, Spain), and under the support of the project PRIN-SURWEY (grant 2012F42NS8, Italy). The authors thank the Editor and the reviewers for their helpful suggestions.

References

Arcos, A., D. Molina, M. Rueda, and M. G. Ranalli (2015). Frames2: A package for estimation in dual frame surveys. *The R Journal*, 7, 52–72.

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.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, 279–292.

Brick, J. M., S. Dipko, S. Presser, C. Tucker, and Y. Yuan (2006). Nonresponse bias in a dual frame survey of cell and landline numbers. *Public Opinion Quarterly*, 70, 780–793.

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

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

Godambe, V. P. and M. E. Thompson (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127–138.

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

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

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

Lehtonen, R. and A. Veijanen (1998). On multinomial logistic generalized regression estimators. Technical Report 22, Department of Statistics, University of Jyväskylä.

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

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

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

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

Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100, 1429–1442.

Ranalli, M., A. Arcos, M. Rueda, and A. Teodoro (2015). Calibration estimation in dual-frame surveys. *Statistical Methods and Applications First online: 01 September 2015*, 1–29.

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

Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Singh, A. C. and F. Mecatti (2011, 12). Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *Journal of official statistics*, 27, 1–19.

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

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

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

Wu, C. and R. R. Sitter (2001a). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185–193.

Wu, C. and R. R. Sitter (2001b). Variance estimation for the finite population distribution function with complete auxiliary information. *Canadian Journal of Statistics*, 29, 289–307.