Estimation techniques for ordinal data in multiple frame surveys with complex sampling designs

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

[1]Department of Statistics and O. R., University of Granada, Spain
E-mail: mrueda@ugr.es
[2]Department of Political Science, University of Perugia, Italy

April 21, 2017

Summary

Surveys usually include questions where individuals must select one in a series of possible options that can be sorted. On the other hand, multiple frame surveys are becoming a widely used method to decrease bias due to undercoverage of the target population. In this work, we propose statistical techniques for handling ordinal data coming from a multiple frame survey using complex sampling designs and auxiliary information. Our aim is to estimate proportions when the variable of interest has ordinal outcomes. Two estimators are constructed following model assisted generalized regression and model calibration techniques. Theoretical properties are investigated for these estimators. Simulation studies with different sampling procedures are considered to evaluate the performance of the proposed estimators in finite size samples. An application to a real survey on opinions towards immigration is also included.

Keywords: complex surveys, generalized regression estimation, model assisted inference, model calibration, multiple frames.

1 Introduction

Dual frame surveys were first introduced by Hartley (1962) as a device for reducing data collection costs without affecting the accuracy of the results with respect to single frame surveys. In general, multiple frame surveys are useful when no single frame covers the whole target population but the union of several (possibly overlapping) available frames does, or when information about a subpopulation of particular interest is obtained only from an

incomplete frame. Independent samples are then selected from each frame, possibly using different sampling designs. Adjustments have to be used at the estimation stage to deal with overlapping and possible unit replication. Multiple frame sampling theory has experienced a noticeable development and several estimators for the total of a quantitative variable have been proposed. In a dual frame context, other than Hartley (1962), classical proposals are in Lund (1968), Fuller & Burmeister (1972), Bankier (1986), Kalton & Anderson (1986) and Skinner (1991). Skinner & Rao (1996) and Rao & Wu (2010) applied likelihood methods to compute estimators that perform well in complex designs. More recently, Ranalli et al. (2016) have used calibration techniques to derive estimators in the dual frame context that make use of auxiliary information. Most of these estimators are implemented in the R package Frames2 (Arcos et al., 2015).

In recent years, a number of works that focus on estimation issues in surveys that use three or more sampling frames has also arisen. Lohr & Rao (2006) extended some of the estimators proposed so far to the multiple frame setting. 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 effective and easy to compute. Multiplicity is also used by Rao & Wu (2010) to provide an extension of the pseudo empirical likelihood estimator to the case of more than two frames. In 2011, Singh & 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.

Surveys in general usually include questions in which the respondents have to indicate their opinion or their degree of agreement with a statement by selecting one of a list of given options. This is the case, particularly, in surveys focused on health, marketing and public

opinion topics. In most situations, the Likert scale is used to scale the possible responses or these are such that they can be ordered according to an intrinsic characteristic of the responses themselves (e.g., from the worst to the best opinion). The main aim is to estimate the proportion of individuals selecting each option. In addition, it is also of interest to estimate the proportion of individuals below (or above) a certain option. This type of variables is also common when using multiple frame surveys and, therefore, estimation techniques should be adjusted to account for the ordinal nature of such variable of interest. In fact, in these situations, classical multiple frame estimators may be used, but the final estimates they provide may suffer from lack of *coherence* or *internal consistency*, in the sense that the sum over the estimates of the proportion of each option may not be equal to one.

The aim of this paper is to propose new estimation techniques for the proportions of variables with ordinal outcomes when data come from multiple frames. In particular, we propose to work within a model assisted framework to finite population inference and make use of auxiliary information to increase the precision of the final estimates. In order to take into account the ordinal nature of the variable of interest, we will use Ordinal Logistic Models (OLMs) to describe the relationship between the variable of interest and the auxiliary variables. This class of models allows to model all the categories of the response variable at the same time and, therefore, provides estimators that are internally consistent, as described above, by definition. In particular, we will introduce estimators based on the model assisted generalized regression estimation approach of Särndal et al. (1992) and on the model calibration approach of Wu & Sitter (2001). Although OLMs have been extensively used in sociological, medical and educational applications, their use for finite population parameter estimation from survey sampling is very sparse.

Molina et al. (2015) also develop estimators for proportions of categorical variables that use auxiliary information. Differently from the approach proposed here, they focus on a dual frame survey context and on categorical variables whose categories can not be ordered. In particular, they propose model assisted estimators that use multinomial logistic regression models. Extending this approach to handle multiple frames is cumbersome. In addition, it may not be adequate because ordinal variables are better modeled using OLMs that account for the intrinsic ordered nature of the response variable.

This article is organized as follows: Section 2 introduces notation and reviews existing approaches for estimation from multiple frame surveys. In Section 3, we illustrate our proposal and discuss the alternative estimators introduced. Asymptotic properties of the proposed estimators are studied in Section 4. The performance of the estimators for finite size samples is investigated through a series of simulation experiments in Section 5. We apply the proposed estimators to a real data set on a dual frame survey on the perception of immigration in a region of Spain in Section 6. Finally, some concluding remarks are provided in Section 7.

2 Notation and estimation for multiple frame ordinal data

We will employ the notation used in Lohr & Rao (2006) and in Mecatti (2007). Let U be a finite population composed of N units labeled from 1 to N, $U = \{1, ..., k, ..., N\}$ and let $A_1, \ldots, A_q, \ldots, A_Q$ be a collection of $Q \geq 2$ overlapping frames of sizes $N_1, \ldots, N_q, \ldots, N_Q$, respectively. All of them can be incomplete but it is assumed that overall they cover the entire target population U. With Q frames, there are up to $2^Q - 1$ distinct domains. Let the index sets K be the subsets of the range of the frame index $q = 1, \ldots, Q$. For every index set $K \subseteq \{1, \ldots, q, \ldots, Q\}$ a domain is defined as the set $D_K = (\cap_{q \in K} A_q) \bigcap (\cap_{q \notin K} A_q^c)$, where

^{c} denotes the complement of a set. That is, D_K is the subset of units that are covered by all the frames A_q, $q \in K$, and by these frames only.

Assume that we collect data from respondents who provide a single choice from a list of ordered alternatives. We code these alternatives as $1, 2, \ldots, m$, with $1 < 2 < \cdots < m$. Therefore, consider an ordinal m-valued survey variable y and we denote by y_k the value observed for the k-th individual of the population. The objective is to estimate the frequency distribution of y in the population U. To estimate this frequency distribution, we define a set 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 population proportion for each i, that is

$$P_i = \frac{1}{N} \sum_{k \in U} z_{ki}, \quad i = 1, 2, \ldots, m. \tag{1}$$

These proportions can be rewritten as follows

$$P_i = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k \in A_q} \frac{z_{ki}}{m_k}, \quad i = 1, 2, \ldots, m, \tag{2}$$

where m_k denotes the number of frames unit k belongs to, i.e. the multiplicity of the k-th unit. Multiplicities m_k are needed in (2) to weight values z_{ki}, so each value z_{ki} is shared by the m_k frames to which unit k belongs to. Otherwise, those units belonging to more than one frame would count more than once in the overall sum. This approach is equivalent to pooling together the Q frames into a single frame that keeps all duplicated units and replaces z_{ki} by z_{ki}/m_k.

Let s_q be a sample drawn from frame A_q under a particular sampling design, independently for $q = 1, \ldots, Q$ and let $\pi_k(q)$, and $\pi_{kl}(q)$ be the first and second order inclusion probabilities under this sampling design, respectively. Let $d_k(q) = 1/\pi_k(q)$ be the sampling weight for unit k in frame A_q. Let n_q be the size of sample s_q. We assume that duplicated

units (i.e. $s_q \cap s_{q'}$, $q \neq q'$) cannot be identified and that this event has a negligible chance to occur. Then we let $s = \bigcup s_q$.

Usually, population level information about auxiliary variables is available in surveys. Let $\boldsymbol{x}_q = (x_{q1}, x_{q2}, \ldots, x_{qp_q})'$ be a set of p_q auxiliary variables observed in the q-th frame, so that the vector $\mathbf{x}_{qk} = (x_{q1k}, x_{q2k}, \ldots, x_{qp_qk})'$ includes the values taken by the variable $\boldsymbol{x}_q$ on unit k in frame A_q. That is, we consider the case of complete auxiliary information. In addition, we consider the more general case in which auxiliary variables may differ in each frame, i.e. $\boldsymbol{x}_q \neq \boldsymbol{x}_r$, for $q, r = 1, \ldots, Q, q \neq r$. For the sample selected from frame A_q, the values of the variables $\{y_k, \mathbf{x}_{qk}\}$ are observed. Equivalently, $\{z_{k1}, \ldots, z_{ki}, \ldots, z_{km}, \mathbf{x}_{qk}\}$ are known.

Lohr & Rao (2006) formulated the multiple frame extension of some of the estimators originally proposed for the dual frame setting, as those proposed by Hartley (1962, 1974) and by Fuller & Burmeister (1972). Although the optimal version of each of these estimators is unbiased and asymptotically efficient, it is not internally consistent when applied to estimate P_i, $i = 1, \ldots, m$, since a different set of weights is used for each dummy variable z_i. Moreover, the multiple frame extension of the Fuller & Burmeister (1972) estimator is often unstable in small to moderate samples, because it requires the estimation of large variance-covariance matrices of estimators.

Lohr & Rao (2006) also followed the *single frame* approach in Kalton & Anderson (1986) to propose a design unbiased estimator in a multiple frame context. In particular, when applied to the issue of estimating P_i, this could be written as

$$\hat{P}_{KAi} = \frac{1}{N} \sum_{k \in s} d_k^{KA} z_{ki} \tag{3}$$

with $d_k^{KA} = (1/\pi_k^+)$, where $\pi_k^+ = \sum_{q:k \in A_q} \pi_k(q)$. In order to compute this estimator, it is necessary to know not only the number of frames each unit belongs to, but also the

specific frames the unit is included in, together with the inclusion probability. This can be an important drawback in multiple frame surveys, particularly if misclassification issues are present.

Mecatti (2007) also considered a single frame approach and proposed the multiplicity adjusted estimator, that can be written here as

$$\hat{P}_{Mi} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k \in s_q} d_k^M(q) z_{ki}, \tag{4}$$

with $d_k^M(q) = d_k(q)/m_k$. This estimator is design unbiased and only requires the knowledge of the multiplicity of each unit, i.e. the number of frames the unit belongs to, no matter which these frames are. Therefore, estimator (4) requires much less information than the estimator in equation (3) but it may be unstable when some units have very small inclusion probabilities. Singh & Mecatti (2011) propose to combine these two estimators. In particular, such composite multiplicity estimator can be written as

$$\hat{P}_{CMi} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{CM}(q) z_{ki}, \tag{5}$$

where

$$d_k^{CM}(q) = \lambda_k d_k^M(q) + (1 - \lambda_k) d_k^{KA}$$

with $\lambda_k \in (0, 1)$. The value of λ_k is obtained minimizing the variance of $d_k^{CM}(q)$ (see Singh and Mecatti, 2001, Appendix A) and is given as a least square type solution by:

$$\lambda_k = \frac{\sum_{q:k \in A_q} (1 - \pi_k^+ / \pi_k(q) m_k) \pi_k(q)(1 - \pi_k(q))}{\sum_{q:k \in A_q} \left(1 - \frac{(\pi_k^+)^2}{\pi_k(q)^2 m_k^2} - \frac{2\pi_k^+}{\pi_k(q) m_k}\right) \pi_k(q)(1 - \pi_k(q))}.$$

Calibration is a well-known technique to exploit auxiliary information in estimation. Ranalli et al. (2016) propose different calibration estimators for the dual frame case, which can be extended to the multiple frame context. A calibration estimator in the case of more

than two sampling frames can be defined as

$$\hat{P}_{CALi} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{CAL}(q) z_{ki}, \tag{6}$$

where $d_k^{CAL}(q)$ are such that they minimize $\sum_{q=1}^{Q} \sum_{k \in sq} G\left(d_k^{CAL}(q), d_k^{M}(q)\right)$ subject to

$$\sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{CAL}(q) \delta_k(A_q) = N_q, \quad q = 1, ..., Q, \tag{7}$$

$$\sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{CAL}(q) \delta_k(A_q) \mathbf{x}_{qk} = \mathbf{t}_{xq}, \quad q = 1, ..., Q. \tag{8}$$

Here $G(\cdot, \cdot)$ is a particular distance function (see Deville & Särndal, 1992, for examples and properties of such functions), $\delta_k(A_q)$ is the indicator variable that takes value 1 if unit k is in frame A_q and zero otherwise, and $\mathbf{t}_{xq}$ are the population totals of $\boldsymbol{x}_q$, $q = 1, \ldots, Q$. Note that weights $d_k^{CAL}(q)$ do not need extra adjustments for multiplicity for two main reasons. First, basic weights $d_k^{M}(q)$ already include the multiplicity adjustment. Therefore, resulting weights $d_k^{CAL}(q)$ should be near to those starting weights that already take into account the multiplicity. Note that, indeed, also other multiplicity adjusted weights as d_k^{KA} or $d_k^{CM}(q)$ could be used as starting weights. Second, and most important, benchmark constraints in (7) and in (8) include all available information coming from the sample on frame A_q, for $q = 1, \ldots, Q$. That is, the indicator variable $\delta_k(A_q)$ takes value 1 for all units belonging to frame A_q, irrespective of the frame they were originally selected from. Therefore, multiplicity is accounted for automatically by the constraints. Note that internal consistency is granted by this type of calibration, because the set of weights $d_k^{CAL}(q)$ is the same for all dummy variables z_i, for $i = 1, \ldots, m$.

It is well-known that calibration, although apparently completely model free, implicitly assumes that a linear regression model well describes the relationship between the variable of interest and the auxiliary variables (see e.g. discussion on this in Wu & Sitter, 2001,

and in Montanari & Ranalli, 2005). In this case the variable of interest is binary, z_i, for $i = 1, \ldots, m$, and this assumption doesn't seem to be adequate. More in general, all the estimators reviewed in this section were originally formulated for estimating parameters (usually a total or a mean) of a quantitative variable. They can be used also for estimating proportions of an ordinal variable although final estimates may likely be unacceptable, in the sense that they can take values outside the interval $[0, 1]$ and they may not add up to one, particularly when different sets of weights are used. Moreover, they are not taking into account the extra information we have from the order among categories. In the following section, we formulate some proposals for estimating proportions of ordinal response variables that address these issues.

3 Proposed estimators

As stated before, we work within the model assisted framework and wish to use ordinal logistic models (OLMs) that are more appropriate for the problem at hand. As it is customary in the model assisted approach to inference (see e.g. Wu & Sitter, 2001, when working with nonlinear and generalized linear models), we first assume that an OLM well describes the relationship between the variable of interest and the auxiliary variables. Then, we obtain parameter estimates for the OLM from sample data using design weighted maximum likelihood techniques, and corresponding predictions for non-sampled units. Finally, we use such predictions in estimators that are inspired by the model assisted generalized regression estimators of Särndal et al. (1992) and of Lehtonen & Veijanen (1998), and in model calibration estimators inspired by those in Wu & Sitter (2001). The proposed estimators are all adjusted for the multiplicity issue that is distinctive of the multiple frame survey framework.

Within OLMs, the most widely used model is the cumulative ordinal logistic model, which assumes a linear model for the logit of cumulative probabilities for the categories of y. See Agresti (2007) for a good introduction to OLMs. Note that, since we consider the most general case, where auxiliary information differs by frame, then we specify a different OLM in each frame. So, in frame A_q, we assume that the relationship between the variable of interest y and the auxiliary variables is well described by the following model, ξ_q,

$$\text{logit}(P(y_k \leq i)) = \log \frac{P(y_k \leq i)}{P(y_k > i)} = \alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}_i^q, \quad i = 1, ..., m-1, \quad q = 1, \ldots, Q, \tag{9}$$

where α_i^q is a scalar and $\boldsymbol{\beta}_i^q = (\beta_{1i}^q, \ldots, \beta_{p_q i}^q)'$ is a vector of parameters. This expression can be rewritten as

$$P(y_k \leq i) = \frac{\exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}_i^q)}{1 + \exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}_i^q)}, \quad i = 1, ..., m-1, \quad q = 1, \ldots, Q. \tag{10}$$

This implies that the (model) expectation of the binary variable z_i is modeled as a function of the auxiliary variables, in fact

$$E_{\xi_q}(z_{ki}|\mathbf{x}_{qk}) = P(y_k = i|\mathbf{x}_{qk}) = \mu_i^q(\mathbf{x}_{qk}),$$

where

$$\mu_i^q(\mathbf{x}_{qk}) = \begin{cases} \dfrac{\exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}_i^q)}{1 + \exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}_i^q)}, & i = 1 \\ \dfrac{\exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}_i^q)}{1 + \exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}_i^q)} - \dfrac{\exp(\alpha_{i-1}^q + \mathbf{x}'_{qk}\boldsymbol{\beta}_{i-1}^q)}{1 + \exp(\alpha_{i-1}^q + \mathbf{x}'_{qk}\boldsymbol{\beta}_{i-1}^q)}, & i = 2, ..., m \end{cases}. \tag{11}$$

Here E_{ξ_q} denotes the expected value with respect to the model in frame A_q.

In proportional odds models, it is assumed that the effects of the predictors are the same across categories. This implies that $\boldsymbol{\beta}_i^q = \boldsymbol{\beta}^q$, i.e. parameters associated to auxiliary variables are common to all the categories considered. This assumption can be tested on sample data.

Then, model (11) can be simplified to

$$\mu_i^q(\mathbf{x}_{qk}) = \begin{cases} \dfrac{\exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}^q)}{1+\exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}^q)}, & i = 1 \\ \dfrac{\exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}^q)}{1+\exp(\alpha_i^q + \mathbf{x}'_{qk}\boldsymbol{\beta}^q)} - \dfrac{\exp(\alpha_{i-1}^q + \mathbf{x}'_{qk}\boldsymbol{\beta}^q)}{1+\exp(\alpha_{i-1}^q + \mathbf{x}'_{qk}\boldsymbol{\beta}^q)}, & i = 2, ..., m \end{cases}. \tag{12}$$

Population parameters α_i^q and $\boldsymbol{\beta}^q$ involved in model ξ_q are unknown and must be estimated using sample information. Different procedures, as weighted least squares (Goldberger, 1964) or weighted maximum likelihood (Binder, 1983), can be used to this end. Under the latter, we can obtain the maximum likelihood estimates for the parameter $\boldsymbol{\theta}^q = (\alpha_1^q, \ldots, \alpha_m^q, \boldsymbol{\beta}^q)$ by maximizing the following function

$$\ell(\boldsymbol{\theta}^q) = \sum_{i=1,...,m} \sum_{k \in s_q} d_k^M(q) z_{ki} \log \mu_i^q(\mathbf{x}_{qk}, \boldsymbol{\theta}^q), \tag{13}$$

and we denote it by $\hat{\boldsymbol{\theta}}^q = (\hat{\alpha}_1^q, \ldots, \hat{\alpha}_m^q, \hat{\boldsymbol{\beta}}^q)$. Under general conditions, the design weighted log-likelihood estimator is design consistent for $\boldsymbol{\theta}^q$ (Binder, 1983; Nordberg, 1989; Wu & Sitter, 2001). It is important to note that, since different auxiliary information is considered in each frame, we need to adjust Q different models, each one based on the set of auxiliary variables of the specific frame.

Using these maximum likelihood estimates, we can define a prediction for probabilities for each category and each unit as follows:

$$p_{ki}^q = \hat{\mu}_i^q(\mathbf{x}_{qk}) = \begin{cases} \dfrac{\exp(\hat{\alpha}_i^q + \mathbf{x}'_{qk}\hat{\boldsymbol{\beta}}^q)}{1+\exp(\hat{\alpha}_i^q + \mathbf{x}'_{qk}\hat{\boldsymbol{\beta}}^q)}, & i = 1 \\ \dfrac{\exp(\hat{\alpha}_i^q + \mathbf{x}'_{qk}\hat{\boldsymbol{\beta}}^q)}{1+\exp(\alpha_i^q + \mathbf{x}'_{qk}\hat{\boldsymbol{\beta}}^q)} - \dfrac{\exp(\hat{\alpha}_{i-1}^q + \mathbf{x}'_{qk}\hat{\boldsymbol{\beta}}^q)}{1+\exp(\hat{\alpha}_{i-1}^q + \mathbf{x}'_{qk}\hat{\boldsymbol{\beta}}^q)}, & i = 2, ..., m \end{cases}. \tag{14}$$

These estimated probabilities can be used to define the following model assisted estimator:

$$\hat{P}_{MAi} = \frac{1}{N}\left[\sum_{q=1}^{Q}\sum_{k \in s_q} z_{ki} d_k^M(q) + \sum_{q=1}^{Q}\left(\sum_{k \in A_q} \frac{p_{ki}^q}{m_k} - \sum_{k \in s_q} p_{ki}^q d_k^M(q)\right)\right], \quad i = 1, ..., m. \tag{15}$$

To formulate this estimator we have adapted the approach used by Lehtonen & Veijanen (1998) to estimate class frequencies of a variable with multinomial outcomes in a single frame

context to the case of an ordinal response variable in a multiple frame setup. Estimated probabilities in the sum over the population are weighted by multiplicities m_k to avoid overestimation issues. For this same reason, weights $d_k^M(q)$ are used in the sample sums. The way in which probabilities are obtained ensures that the estimator is internally consistent in the sense that its categories add up to one, that is $\sum_{i=1}^{m} \hat{P}_{MAi} = 1$.

An Hajek-type estimator can also be constructed by replacing N in (15) by an estimate, e.g. by $\hat{N} = \sum_{q=1}^{Q} \sum_{k \in s_q} d_k^M(q)$. This is a special case of ratio estimator, and it can be more efficient than Horvitz-Thompson type estimators because the sample size in overlapping domains is not fixed.

Treating probabilities p_{ki}^q as auxiliary variables, we can include them in the estimation process through a model calibration approach (Wu & Sitter, 2001). The resulting model calibration estimator can be written as

$$\hat{P}_{MCi} = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{MC}(q) z_{ki}, \quad i = 1, ..., m, \tag{16}$$

where weights $d_k^{MC}(q)$ are chosen so that they minimize $\sum_{q=1}^{Q} \sum_{k \in s_q} G\left(d_k^{MC}(q), d_k^M(q)\right)$, subject to

$$\sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{MC}(q) \delta_k(A_q) = N_q, \quad q = 1, ..., Q$$

$$\sum_{q=1}^{Q} \sum_{k \in s_q} d_k^{MC}(q) p_{ki}^q \delta_k(A_q) = \sum_{q=1}^{Q} \sum_{k \in A_q} p_{ki}^q \delta_k(A_q), \quad q = 1, ..., Q, \quad i = 1, ..., m. \tag{17}$$

Note that $\sum_{q=1}^{Q} \sum_{k \in A_q} p_{ki}^q \delta_k(A_q) = \sum_{k \in A_q} p_{ki}^q$ in (17), and this is in line with the reasoning we had used in $\hat{P}_{CALi}$ in (6). Then, similarly to $\hat{P}_{CALi}$, the proposed model calibration estimator eliminates overestimation issues by several means. First, we consider $d_k^M(q)$ (which are already weighted by m_k) as the starting weights for the calibration. More importantly,

using the indicator variables $\delta_k(A_q)$, the calibration constraints ensure adjustment of the multiplicity issues by benchmarking all information on units from frame A_q included in the sample, irrespective of the frame they were originally selected from. Therefore, again, multiplicity is accounted for automatically by the constraints. Differently from $\hat{P}_{CALi}$, constraints (17) do not implicitly assume a linear model between the auxiliary variables and the response indicators, but properly account for the ordinal nature of the response variable using predictions from the OLM.

Note that to compute the proposed estimators we need to estimate the probabilities p_{ki}^q for each individual in each frame. This implies the knowledge of the auxiliary information for each of these individuals. Although this assumption can be quite restrictive, it can be relaxed in many situations. For example, when qualitative 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, we only need to know the frequency of each possible combination of the values of these auxiliary variables in each frame to compute the proposed estimators. This information can usually be found in the databases of statistical agencies.

4 Properties of the proposed estimators

In this section we describe the main properties of the proposed estimators. We adapt the asymptotic framework of Isaki & Fuller (1982) to a multiple frame context. Such framework has been also used in a dual frame context by Rao & Wu (2010) and by Ranalli et al. (2016). In particular, the finite population U and the sampling designs $p_1(\cdot), p_2(\cdot), ..., p_Q(\cdot)$ are embedded into a sequence of such populations and designs indexed by N, $\{U_N, p_{1_N}(\cdot), p_{2_N}(\cdot), ...,$

$p_{Q_N}(\cdot)\}$, with $N \to \infty$. We will assume, thus, that $N_{1_N}, N_{2_N}, ..., N_{Q_N}$ tend to infinity and that $n_{1_N}, n_{2_N}, ..., n_{Q_N}$ also tend to infinity when $N \to \infty$. Furthermore, we will assume $n_{q_N}/n_N \to c_q \in (0,1), q = 1, \ldots, Q$, where $n_N = \sum_{q=1}^{Q} n_{q_N}$ as $N \to \infty$. All limiting processes are understood as $N \to \infty$, so we drop subscript N for ease of notation. Stochastic orders $O_p(\cdot)$ and $o_p(\cdot)$ are with respect to the aforementioned sequences of designs. We first discuss the theoretical properties of $\hat{P}_{MCi}$ and then move to those of $\hat{P}_{MAi}$, because the latter can be seen as a particular case of the former.

Let $\tilde{\boldsymbol{\theta}}^q$ be the solution to the census level likelihood, that is

$$\ell_U(\boldsymbol{\theta}^q) = \sum_{i=1,\ldots,m} \sum_{k \in A_q} z_{ki} \log \mu_i^q(\mathbf{x}_{qk}, \boldsymbol{\theta}^q), \tag{18}$$

and $\tilde{p}_{ki}^q = \mu_i^q(\mathbf{x}_{qk}, \tilde{\boldsymbol{\theta}}^q)$ for $i = 1, ..., m$ and $q = 1, ..., Q$. In addition, let

$$\boldsymbol{\omega}_k = (\delta_k(A_1), \ldots, \delta_k(A_Q), \delta_k(A_1)p_{k1}^1, \ldots, \delta_k(A_Q)p_{k1}^Q, \ldots, \delta_k(A_1)p_{km}^1, \ldots, \delta_k(A_Q)p_{km}^Q)'$$

be the $Q + Q \times m$ vector of all auxiliary variables used in the benchmarking constraints for $\hat{P}_{MCi}$, and let

$$\tilde{\boldsymbol{\omega}}_k = (\delta_k(A_1), \ldots, \delta_k(A_Q), \delta_k(A_1)\tilde{p}_{k1}^1, \ldots, \delta_k(A_Q)\tilde{p}_{k1}^Q, \ldots, \delta_k(A_1)\tilde{p}_{km}^1, \ldots, \delta_k(A_Q)\tilde{p}_{km}^Q)'$$

be its population level counterpart. In order to prove our results, we make a set of technical assumptions reported in Appendix A.1.

Theorem 4.1. *Under assumptions A1–A3, estimator $\hat{P}_{MCi}$ is design $\sqrt{n_N}$-consistent for P_i in the sense that*

$$\hat{P}_{MCi} - P_i = O_p(n_N^{-1/2}),$$

and has the following asymptotic distribution

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

where

$$V_\infty(\hat{P}_{MCi}) = V\left(\frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in s_q} d_k^M(q)e_{ki}\right) = \frac{1}{N^2}\sum_{q=1}^{Q} V\left(\sum_{k\in s_q} d_k(q)\frac{e_{ki}}{m_k}\right) \tag{19}$$

with population level residuals $e_{ki} = z_{ki} - \tilde{\boldsymbol{\omega}}_k'\tilde{\boldsymbol{\gamma}}_i$ *and*

$$\tilde{\boldsymbol{\gamma}}_i = \left(\sum_{k\in U}\tilde{\boldsymbol{\omega}}_k\tilde{\boldsymbol{\omega}}_k'\right)^{-1}\sum_{k\in U}\tilde{\boldsymbol{\omega}}_k z_{ki}. \tag{20}$$

In addition, under assumptions A1–A5, $V_\infty(\hat{P}_{MCi})$ *can be consistently estimated by*

$$v(\hat{P}_{MCi}) = \frac{1}{N^2}\sum_{q=1}^{Q} v\left(\sum_{k\in s_q} d_k(q)\frac{\hat{e}_{ki}}{m_k}\right),$$

where $v(\cdot)$ *is the Horvitz-Thompson variance estimator of* $V(\cdot)$ *with* $\hat{e}_{ki} = z_{ki} - \boldsymbol{\omega}_k'\hat{\boldsymbol{\gamma}}_i$ *and*

$$\hat{\boldsymbol{\gamma}}_i = \left(\sum_{q=1}^{Q}\sum_{k\in s_q} d_k^M(q)\boldsymbol{\omega}_k\boldsymbol{\omega}_k'\right)^{-1}\sum_{q=1}^{Q}\sum_{k\in s_q} d_k^M(q)\boldsymbol{\omega}_k z_{ki}. \tag{21}$$

Proof. See Appendix A.2

■

Estimator $\hat{P}_{MAi}$ can be seen as a particular case of estimator $\hat{P}_{MCi}$. This is a common finding when comparing model assisted generalized regression type estimators with model calibration estimators (see e.g. Wu & Sitter, 2001; Montanari & Ranalli, 2005). In fact, $\hat{P}_{MAi}$ uses only one auxiliary variable given by p_{ki}^q, for $q = 1, \ldots, Q$ and is equivalent to $\hat{P}_{MCi}$ as in equation (23) if we use p_{ki}^q as auxiliary variable with benchmark constraint given by $\sum_q \sum_{A_q} p_{ki}^q/m_k$, and set $\hat{\gamma} = 1$. Therefore, we can summarize properties of $\hat{P}_{MAi}$ in the following Theorem. The proof is immediate and is omitted.

Theorem 4.2. *Under assumptions A2–A3, estimator* $\hat{P}_{MAi}$ *is design* $\sqrt{n_N}$*-consistent for* P_i *in the sense that*

$$\hat{P}_{MAi} - P_i = O_p(n_N^{-1/2}),$$

and has the following asymptotic distribution

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

where

$$V_\infty(\hat{P}_{MAi}) = V\left(\frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in s_q} d_k^M(q) e_{ki}^q\right) = \frac{1}{N^2}\sum_{q=1}^{Q} V\left(\sum_{k\in s_q} d_k(q)\frac{e_{ki}^q}{m_k}\right) \tag{22}$$

with population level residuals $e_{ki}^q = z_{ki} - \tilde{p}_{ki}^q$.

In addition, under assumptions A2–A5, $V_\infty(\hat{P}_{MAi})$ *can be consistently estimated by*

$$v(\hat{P}_{MAi}) = \frac{1}{N^2}\sum_{q=1}^{Q} v\left(\sum_{k\in s_q} d_k(q)\frac{\hat{e}_{ki}^q}{m_k}\right),$$

where $v(\cdot)$ *is the Horvitz-Thompson variance estimator of* $V(\cdot)$ *with* $\hat{e}_{ki}^q = z_{ki} - p_{ki}^q$.

5 Monte Carlo Simulation Experiments

We now present the results of some Monte Carlo experiments carried out to empirically compare the performance of the proposed estimators with respect to the customary estimators discussed in Section 2. To carry out the simulation study we have used the freeware statistical program R.

We have considered a three frame setting, frames A_1, A_2 and A_3, where three normal variables have been simulated: a first one following a $\mathcal{N}(30, 3)$, which is categorized considering 4 ordered levels to create the ordinal response variable, y, (for simplicity, we have coded the levels as 1, 2, 3 and 4, considering $1 < 2 < 3 < 4$) and another two which play the role of auxiliary variables: x_1 and x_2. These two auxiliary variables are generated controlling their correlation with the response variable (taking advantage of the fact that response variable has been generated from a continuous variable). In this first scenario, the correlation between the continuous variable behind the response y and the auxiliary variables x_1 and

x_2 has been set at 0.85. We have generated $N = 10000$ observations for each of the three variables involved in the study. Population proportions for the levels of the ordinal response variable are: 0.1, 0.2, 0.3 and 0.4, respectively.

Domain sizes were defined beforehand and then each unit was randomly assigned to one of the domains. As a result, three overlapping frames of sizes $N_1 = 5500, N_2 = 6000$ and $N_3 = 5000$ were obtained. Three samples of sizes $n_1 = 360, n_2 = 464$ and $n_3 = 728$ were independently drawn, one from each frame, considering Midzuno sampling designs in frames A_1 and A_3 and a simple random sampling design in frame A_2. Sample from frame A_1 was drawn with probabilities proportional to a normally distributed variable with mean 1000 and standard deviation 250. On the other hand, sample from frame A_3 was drawn considering inclusion probabilities proportional to another normally distributed variables with mean 5000 and standard deviation 500. These two normal variables are such that the correlation between each of them and the continuous variable behind the response y is 0.9. In this scenario, the ordinal model-assisted estimator (PMA) and the ordinal model-calibrated estimator (PMC) were computed. For comparison purposes, we also compute Kalton-Anderson (KA), multiplicity (M), composite multiplicity (CM) and calibration (CAL) estimators. For the estimators using auxiliary information (CAL, PMA and PMC) we have considered different sets of variables: x_1 in frame A_1, x_2 in frame A_2 and both x_1 and x_2 in frame A_3.

For each estimator, we compute the percent relative bias $RB\% = E_{MC}(\hat{P} - P)/P * 100$ and the percent relative mean squared error $RMSE\% = E_{MC}[(\hat{P} - P)^2]/P^2 * 100$ for each category of the variable y based on 10000 simulation runs. We have used $RMSE\%$ to calculate percent relative efficiency gain with respect to multiplicity estimator. This percent relative efficiency gain for a generic estimator $\hat{P}$ is defined as $RMSE\%_M/RMSE\%_{\hat{P}} * 100$,

where $RMSE\%_M$ is the percent relative mean squared error for the multiplicity estimator (results are presented in Table 1).

Table 1: % Relative bias (in italics) and % relative efficiency, with respect to multiplicity estimator for each estimator. Corresponding equation in parentheses. $\rho_{YX_1} = 0.85$, $\rho_{YX_2} = 0.85$

	1	2	3	4	min	max	mean
M (4)	*0.00*	*0.01*	*0.07*	*-0.05*	*0.00*	*0.07*	*0.03*
	100.00	100.00	100.00	100.00	100.00	100.00	100.00
KA (3)	*0.01*	*0.01*	*0.07*	*-0.06*	*0.01*	*0.07*	*0.04*
	103.54	104.02	104.44	103.73	103.54	104.44	103.93
CM (5)	*0.02*	*0.02*	*0.06*	*-0.05*	*0.01*	*0.08*	*0.03*
	103.61	104.39	104.16	104.38	103.61	104.38	104.13
CAL (6)	*-0.34*	*0.05*	*0.22*	*-0.11*	*0.05*	*0.34*	*0.18*
	129.76	115.42	99.66	171.85	99.66	171.85	129.17
PMA (15)	*0.68*	*-0.26*	*-0.23*	*0.13*	*0.13*	*0.68*	*0.32*
	181.01	135.62	124.38	212.51	124.38	212.51	163.38
PMC (16)	*-0.11*	*-0.02*	*0.06*	*-0.01*	*0.01*	*0.11*	*0.05*
	174.40	133.01	123.63	192.83	123.63	192.83	155.96

From results of Table 1 we can conclude that bias for all the estimators considered is negligible. Equally, we can observe that estimators using auxiliary variables perform better than the estimators that do not use any extra information. The proposed ordinal estimators work better than the classical calibration estimator, which assume an underlying linear model. Whatever the proposed estimator, we can see that the largest mean efficiency gain with respect to multiplicity estimator is achieved in category 4, which is the category with the largest population proportion. The PMA estimator shows a slightly better performance than the PMC estimator in terms of efficiency gain.

To determine the effect of varying association between the main variable and auxiliary variables, we are going to consider new scenarios with different correlation levels between the

continuous variable behind the response y and x_1 and x_2. In the first scenario, correlation between the continuous variable behind the response y and x_1 has been decreased with respect to the initial situation to 0.65 and correlation between the continuous variable behind the response y and x_2 has been set to 0.5. In the second scenario, correlation levels between the continuous variable behind the response y and x_1 and between the continuous variable behind the response y and x_2 are set to 0.4 and 0.7, respectively. We have run 10000 replications keeping the same sample sizes for the three frames. Relative bias is not significant in any case and so only relative efficiency with respect to multiplicity estimator is displayed in Table 2.

Table 2: % Relative efficiency with respect to multiplicity estimator of compared estimators considering different association levels between y and x_1 and x_2

	1	2	3	4	min	max	mean
			$\rho_{YX_1} = 0.65$, $\rho_{YX_2} = 0.5$.				
PMA	121.19	110.37	105.25	131.59	105.25	131.59	117.10
PMC	121.46	108.71	103.85	130.76	103.85	130.76	116.19
			$\rho_{YX_1} = 0.4$, $\rho_{YX_2} = 0.7$.				
PMA	122.04	110.64	106.17	131.30	106.17	131.30	117.53
PMC	121.62	109.19	104.71	131.57	104.71	131.57	116.77

We observe that the proposed estimators have a gain in efficiency in comparison to the customary multiplicity estimator when the association between the auxiliary variables and the main variable is also moderated. If correlation decreases, then the improvement of course of using the model is less important. As in the previous scenario, gain in efficiency for category 4 is quite relevant compared with the 3 remaining categories.

We have also computed confidence intervals considering two different approaches for estimating the variance of the proposed estimators: the jackknife procedure described in Rao

& Wu (2010) and the analytic expression for the estimators of the variances formulated in Section 4. Table 3 shows the length reduction of 95% confidence intervals with respect to the multiplicity estimator and the empirical coverage probability over 10000 simulation runs in each category of the main variable. We can see that the proposed estimators considerably reduce the length of the confidence intervals obtained, with respect to the multiplicity estimator irrespective of the method used for variance estimation. More importantly, the empirical coverage is always very close to the nominal level.

Table 3: % length reduction with respect to multiplicity estimator (in italics) and empirical coverage of 95% confidence intervals for the estimators using the jackknife method and the analytic expression for the variance estimation.

	Jackknife				
	1	2	3	4	mean
PMA	*25.82*	*13.46*	*10.66*	*31.09*	*20.26*
	96.08	96.20	96.64	96.20	96.28
PMC	*22.88*	*12.29*	*9.69*	*26.19*	*17.76*
	94.97	96.08	96.20	95.53	95.69
	Analytic				
	1	2	3	4	mean
PMA	*26.51*	*13.53*	*10.15*	*30.85*	*20.26*
	95.48	96.09	97.17	96.31	96.26
PMC	*26.51*	*13.44*	*10.06*	*30.77*	*20.19*
	94.68	96.60	96.44	95.39	95.78

6 Application to real data

In this Section we report on the results of application of the proposed estimators to real data from an opinion survey. In particular, data come from a survey on opinions of the Andalusian population towards immigration conducted in 2013 by an Andalusian research institute focusing on social studies. In this survey, the institute conducting the survey decided

to carry out telephone interviews with adults using two sampling frames: one of landlines (frame A_1) and another one of cell phones (frame A_2). Overall, $n = 1853$ telephone interviews were performed.

At the time of data collection, frame sizes were known (extracted from ICT-H 2012, Survey on the Equipment and Use of Information and Communication Technologies in Households, INE, National Statistical Institute, Spain). Landline frame was stratified by provinces in the region of Andalusia and then a stratified sample of size $n_1 = 1468$ was drawn. In the cell phone frame a simple random sample of size $n_2 = 385$ was selected by using random digit dialing.

We have considered the response to the question "In relation to the number of immigrants currently living in Andalusia, do you think there are too many, a reasonable number or too few?" as main variable of interest. As auxiliary information we have used the age (categorized into 4 age classes) of interviewed people in each frame. We have tested for the proportional odds assumption in the data: the p-values associated to the test are 0.1492 and 0.0725 in frame A and in frame B, respectively. Population data for auxiliary variables is reported in Table 4.

Together with the proposed estimators, we have calculated some additional estimators for comparison purposes as the multiplicity (M), Kalton-Anderson (KA), composite multiplicity (CM) and calibration (CAL) estimators. For CAL estimator we have used also the age of the individuals as auxiliary variable.

Table 5 shows point estimation for the considered estimators for the main variable. We have used the jackknife procedure described in Rao & Wu (2010) as well as the analytic expression for the estimator of the variance to compute a 95 % confidence interval. Results

Table 4: Population data for variable `age`

	Both	Landline	Cell
18 - 29	908,901	0	303,644
30 - 44	1,383,419	21,932	587,522
45 - 59	1,204,816	98,747	277,534
> 60	842,523	522,582	199,296

of lower bound, upper bound and length of confidence intervals for each method of variance estimation are also included in the table.

In both cases, average length of confidence intervals of all proposed estimators is smaller than average lengths of confidence intervals of classical estimators. This fact can be seen in Table 6 that shows the reduction of the length of the proposed estimator respect to the multiplicity estimator.

7 Conclusions

In this paper we have introduced a flexible way of using auxiliary information when estimating proportions for an ordinal variable using a multiple frame survey. We have worked within the model-assisted framework for finite population inference and proposed estimators using both the generalized regression and the calibration approach. In both cases, we have relaxed the assumption of a linear regression model and considered ordinal regression models. Weighted likelihood methods have been employed to obtain design consistent parameter estimates. The properties of the proposed estimators have been investigated theoretically and via simulation studies.

The performance of the proposed ordinal estimators is good under a variety of sampling designs. Our main findings show that it is important to include auxiliary information into the estimation process to increase efficiency. Of course, the gain in efficiency depends on

the strength of the relationship of the auxiliary variables with the variable of interest. In addition, it is also important to account for the ordinal nature of the variable of interest and, therefore, employ suitable assisting models. In fact, the proposed estimators outperform classical calibration methods that, implicitly, employ a linear regression model. In this regard, a methodology that is often used to incorporate auxiliary information in sample surveys is post-stratification; it should be noted that it is just a particular case of calibration and, therefore, we have shown that it is possible to use auxiliary information in a more efficient way when the variable of interest is ordinal. This has been highlighted also in the application to real data from a dual frame survey on attitudes towards immigration: the calibration estimator in this case is essentially an adaptation of post-stratification to multiple frame surveys. The proposed estimators provide all a sensible reduction on the length on the confidence intervals for the estimated proportions compared to all other estimators.

Acknowledgements

This study was partially supported by Ministerio de Economía y Competitividad (grant MTM2015-63609-R 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).

References

[1] Agresti, A. (2007). An introduction to categorical data analysis, 2nd ed. *Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.*

[2] Arcos, A., Molina, D., Rueda, M. and Ranalli, M. G. (2015). Frames2: A package for estimation in dual frame surveys. *R J.*, **7**, 52-72.

[3] Bankier, M. D. (1986). Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys. *J. Amer. Statist. Assoc.*, **81**, 1074–79.

[4] Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Int. Stat. Rev.* **51**(3), 279–292.

[5] Breidt, F. J., Claeskens, G. and Opsomer, J. D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, **92**, 831–846.

[6] Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling *J. Amer. Statist. Assoc.*, **87**, 376 – 382.

[7] Fuller, W. A. and Burmeister, L. F. (1972). Estimators for samples selected from two overlapping frames. In *Proceedings of the American Statistical Association, Social Statistics Section*, 245–249.

[8] Goldberger, A. S. *Econometric theory.* New York: Wiley, 1964

[9] Hartley, H. O. (1962). Multiple Frame Surveys. In *Proceedings of the American Statistical Association, Social Statistics Sections*, 203–206.

[10] Hartley, H. O. (1974). Multiple frame methodology and selected applications. *Sankhya Ser. C*, **36**, 99 – 118.

[11] Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model *J. Amer. Statist. Assoc.*, **77**(377), 89–96.

[12] Kalton, G. and Anderson, D. W. (1986). Sampling rare populations. *J. Roy. Statist. Soc. Ser. A*, **149**, 65–82.

[13] Lehtonen, R., and A. Veijanen. (1998). Logistic generalized regression estimators. *Surv. Methodol.* **24**, 51–55.

[14] Lohr, S.: Multiple frame surveys. In: Rao, C.R., Pfeffermann, D. (eds.) Handbook of

Statis- tics, Vol. 29A, Sample Surveys: Design, Methods and Applica tions, pp. 71–88. North Holland, Amsterdam (2009)

[15] Lohr, S. and Rao, J. N. K. (2006). Estimation in multiple frame surveys. *J. Amer. Statist. Assoc.*, **101**(475), 1019 – 1030.

[16] Lund, R. E. (1968). Estimators in multiple frame surveys. In: Proceedings of the American Statistical Association, Social Science, pp. 282 – 286.

[17] Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Surv. Methodol.*, **33**(2), 151 – 157.

[18] Molina, D., Rueda, M., Arcos, A. and Ranalli, M. (2015). Multinomial logistic estimation in dual frame surveys. *SORT*, **39**(2), 309–336.

[19] Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *J. Amer. Statist. Assoc.*, **472**(100), 1429–1442.

[20] Nordberg, L. (1989). Generalized Linear Modeling of Sample Survey Data. *J. Off. Statis.* **5**, 223-239

[21] Ranalli, M. G., Arcos, A., Rueda, M. and Teodoro, A. (2016). Calibration estimation in dual-frame surveys. *Stat. Methods Appl.*, **25**(3), 321-349

[22] Rao, J. N. K. and Wu, C. (2010). Pseudo empirical likelihood inference for multiple frame surveys. *J. Amer. Statist. Assoc.*, **105**, 1494 – 1503.

[23] Särndal C. E., Swenson B, Wretman J. *Model assisted survey sampling.* New York: Springer-Verlag. 1992.

[24] Singh, A. C. and Mecatti, F. (2011). Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *J. Off. Statis.*, **27**(4), 1 – 19.

[25] Skinner, C. J. (1991). On the efficiency of raking ration estimation for multiple frame surveys. *J. Amer. Statist. Assoc.*, **86**, 779–784.

[26] Skinner, C. J. and Rao, J. N. K. (1996). Estimation in dual frame surveys with complex designs. *J. Amer. Statist. Assoc.* **91**(433), 349–56.

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

Table 5: Point and 95% confidence level estimation of percentages using Jackknife and Analytic variance estimation.

In relation to the number of immigrants currently living in Andalusia, do you think there are . . . ?							
		Jackknife			Analytic		
Estimator	PROP	LB	UB	LEN	LB	UB	LEN
	...too many						
M	45.26	41.93	48.58	6.65	41.09	49.41	8.31
KA	44.92	41.84	48.00	6.17	41.43	48.41	6.97
CM	44.95	41.86	48.05	6.19	41.68	48.23	6.55
CAL	44.68	40.83	48.54	7.71	40.24	49.13	8.89
PMA	45.11	42.10	48.11	6.00	41.93	48.29	6.36
PMC	44.65	41.71	47.59	5.88	41.67	47.57	5.90
	...a reasonable number						
M	48.36	45.03	51.67	6.64	44.29	52.41	8.12
KA	48.64	45.56	51.72	6.16	45.18	52.10	6.91
CM	48.61	45.52	51.69	6.17	45.45	51.76	6.31
CAL	49.26	44.78	53.73	8.95	44.83	53.69	8.86
PMA	48.69	45.69	51.70	6.00	45.58	51.80	6.22
PMC	49.27	46.21	52.32	6.11	46.23	52.38	6.14
	...too few						
M	6.39	4.68	8.09	3.41	4.63	8.13	3.49
KA	6.43	4.83	8.03	3.20	4.80	8.05	3.24
CM	6.43	4.82	8.03	3.21	4.88	7.97	3.09
CAL	6.04	4.25	7.84	3.59	3.86	8.23	4.37
PMA	6.18	4.68	7.69	3.00	4.65	7.71	3.06
PMC	6.07	4.67	7.48	2.81	4.56	7.56	3.00

Table 6: Relative length reduction in % of the 95% confidence intervals of the proposed estimators with respect to the multiplicity estimator.

In relation to the number of immigrants currently living in Andalusia, do you think there are...?		
Estimator	Jackknife	Analytic
	...too many	
PMA	9.66	23.46
PMC	11.65	29.00
	...a reasonable number	
PMA	9.56	23.39
PMC	8.01	24.38
	...too few	
PMA	11.87	12.32
PMC	17.49	14.04

A Appendix - Assumptions and proof of Theorem 4.1.

A.1 Assumptions

A1. $\boldsymbol{\gamma}_i = \lim_{N\to\infty} \tilde{\boldsymbol{\gamma}}_i$ *exists and* $\hat{\boldsymbol{\gamma}}_i = \tilde{\boldsymbol{\gamma}}_i + o_p(1)$, *for* $i = 1, \ldots, m$, *where* $\tilde{\boldsymbol{\gamma}}_i$ *and* $\hat{\boldsymbol{\gamma}}_i$ *are given in (20) and (21), respectively.*

A2. *The limiting design covariance matrix of the normalized, multiplicity adjusted Horvitz-Thompson estimators,*

$$\begin{aligned}\boldsymbol{\Sigma}^q &= \begin{bmatrix} \Sigma^q_{zz} & \boldsymbol{\Sigma}^q_{z\tilde{\omega}} \\ \boldsymbol{\Sigma}^{q\prime}_{z\tilde{\omega}} & \boldsymbol{\Sigma}^q_{\tilde{\omega}\tilde{\omega}} \end{bmatrix} \\ &= \lim_{N\to\infty} \frac{n_N}{N^2} \begin{bmatrix} \sum\sum_{A_q} \Delta_{kl}(q) z_{ki} z_{li} & \sum\sum_{A_q} \Delta_{kl}(q) z_{ki} \tilde{\boldsymbol{\omega}}'_l \\ \sum\sum_{A_q} \Delta_{kl}(q) \tilde{\boldsymbol{\omega}}_k z_{li} & \sum\sum_{A_q} \Delta_{kl}(q) \tilde{\boldsymbol{\omega}}_k \tilde{\boldsymbol{\omega}}'_l \end{bmatrix}\end{aligned}$$

is positive defined, with $\Delta_{kl}(q) = (\pi_{kl}(q) - \pi_k(q)\pi_l(q))/\pi_k(q)\pi_l(q)m_k m_l$, *for* $q = 1, \ldots, Q$.

A3. *The normalized multiplicity adjusted Horvitz-Thompson estimators satisfy a central limit theorem:*

$$\frac{\sqrt{n_N}}{N} \begin{bmatrix} \sum_q \sum_{k\in s_q} d^M_k(q) z_{ki} - P_i \\ \sum_q \sum_{k\in s_q} d^M_k(q) \tilde{\boldsymbol{\omega}}_k - \sum_q \sum_{k\in A_q} \tilde{\boldsymbol{\omega}}_k \end{bmatrix} \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \sum_q \boldsymbol{\Sigma}^q\right).$$

A4. $N^{-1}\left[\sum_q \sum_{k\in s_q} d^M_k(q)(\boldsymbol{\omega}_k - \tilde{\boldsymbol{\omega}}_k) - \sum_q \sum_{k\in A_q} (\boldsymbol{\omega}_k - \tilde{\boldsymbol{\omega}}_k)\right] = o_p(n_N^{-1/2})$.

A5. *The estimated covariance matrix for the Horvitz-Thompson estimators*

$$\hat{\boldsymbol{\Sigma}}^q = \begin{bmatrix} \sum\sum_{s_q} \pi_{kl}(q)^{-1} \Delta_{kl}(q) z_{ki} z_{li} & \sum\sum_{s_q} \pi_{kl}(q)^{-1} \Delta_{kl}(q) z_{ki} \tilde{\boldsymbol{\omega}}'_l \\ \sum\sum_{s_q} \pi_{kl}(q)^{-1} \Delta_{kl}(q) \tilde{\boldsymbol{\omega}}_k z_{li} & \sum\sum_{s_q} \pi_{kl}(q)^{-1} \Delta_{kl}(q) \tilde{\boldsymbol{\omega}}_k \tilde{\boldsymbol{\omega}}'_l \end{bmatrix}$$

is design consistent in the following sense:

$$\frac{n_N}{N^2} \hat{\boldsymbol{\Sigma}}^q - \boldsymbol{\Sigma}^q = o_p(1),$$

for $q = 1, \ldots, Q$.

These assumptions are similar to those used in Breidt, Claeskens and Opsomer (2005) and in Ranalli et al. (2016). Assumption A1 ensures that the sample fit $\hat{\boldsymbol{\gamma}}_i$ and the population fit $\tilde{\boldsymbol{\gamma}}_i$ share a common limit. This assumption, together with Assumption A4, depend on the distribution of the auxiliary variables $\boldsymbol{x}_q$, of the function $\mu(\cdot)$ and parameter estimates of $\boldsymbol{\theta}^q$. Wu & Sitter (2001) provide conditions for the case of generalized linear models. Assumptions A2 and A3 are satisfied for commonly used fixed sample size designs in reasonably finite populations. Assumption A5 is satisfied by many common designs. However, it would not hold for systematic sampling or one-per-stratum designs.

A.2 Proof of Theorem 4.1.

Without loss of generality, we consider the chi-squared distance measure $G(w;d) = (w - d)^2/d^2$. It can be easily shown that in this case $\hat{P}_{MCi}$ can be written as follows

$$\hat{P}_{MCi} = \frac{1}{N}\left[\sum_{q=1}^{Q}\sum_{k\in s_q} d_k^M(q) z_{ki} + \sum_{q=1}^{Q}\left(\sum_{k\in A_q}\boldsymbol{\omega}_k - \sum_{k\in s_q} d_k^M(q)\boldsymbol{\omega}_k\right)' \hat{\boldsymbol{\gamma}}_i\right]. \tag{23}$$

Now,

$$\begin{aligned}
\hat{P}_{MCi} - P_i &= \tilde{P}_{MCi} - P_i + \frac{1}{N}\sum_{q=1}^{Q}\left(\sum_{k\in A_q}\boldsymbol{\omega}_k - \sum_{k\in s_q} d_k^M(q)\boldsymbol{\omega}_k\right)'(\hat{\boldsymbol{\gamma}}_i - \tilde{\boldsymbol{\gamma}}_i) + \\
&\quad + \frac{1}{N}\sum_{q=1}^{Q}\left(\sum_{k\in A_q}(\boldsymbol{\omega}_k - \tilde{\boldsymbol{\omega}}_k) - \sum_{k\in s_q} d_k^M(q)(\boldsymbol{\omega}_k - \tilde{\boldsymbol{\omega}}_k)\right)' \tilde{\boldsymbol{\gamma}}_i \\
&= \tilde{P}_{MCi} - P_i + O_p(n_N^{-1/2})o_p(1) + o_p(n_N^{-1/2})O_p(1),
\end{aligned}$$

where

$$\tilde{P}_{MCi} = \frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in s_q} d_k^M(q) z_{ki} + \frac{1}{N}\sum_{q=1}^{Q}\left(\sum_{k\in A_q}\tilde{\boldsymbol{\omega}}_k - \sum_{k\in s_q} d_k^M(q)\tilde{\boldsymbol{\omega}}_k\right)' \tilde{\boldsymbol{\gamma}}_i. \tag{24}$$

Therefore the asymptotic distribution of $\hat{P}_{MCi}$ is the same as that of $\tilde{P}_{MCi}$, whose variance is $V_\infty(\hat{P}_{MCi})$.

To prove consistency of the variance estimator, note that

$$\begin{aligned} v\left(\sum_{k\in s_q} d_k(q)\frac{\hat{e}_{ki}}{m_k}\right) &= \sum\sum_{s_q} \pi_{kl}(q)^{-1}\Delta_{kl}(q)\hat{e}_{ki}\hat{e}_{li} \\ &= \sum\sum_{s_q} \pi_{kl}(q)^{-1}\Delta_{kl}(q)e_{ki}e_{li} + \\ &+ (\tilde{\boldsymbol{\gamma}}_i - \hat{\boldsymbol{\gamma}}_i)' \sum\sum_{s_q} \pi_{kl}(q)^{-1}\Delta_{kl}(q)\boldsymbol{\omega}_k\boldsymbol{\omega}_l'(\tilde{\boldsymbol{\gamma}}_i - \hat{\boldsymbol{\gamma}}_i) + \\ &+ \tilde{\boldsymbol{\gamma}}_i{}' \sum\sum_{s_q} \pi_{kl}(q)^{-1}\Delta_{kl}(q)(\tilde{\boldsymbol{\omega}}_k - \boldsymbol{\omega}_k)(\tilde{\boldsymbol{\omega}}_l - \boldsymbol{\omega}_l)'\,\tilde{\boldsymbol{\gamma}}_i. \end{aligned}$$

Therefore, $v\left(\sum_{k\in s_q} d_k(q)\frac{\hat{e}_{ki}}{m_k}\right) = v\left(\sum_{k\in s_q} d_k(q)\frac{e_{ki}}{m_k}\right) + N^2 o_p(n_N^{-1})$ by Assumptions A1-A5, and $v\left(\sum_{k\in s_q} d_k(q)\frac{e_{ki}}{m_k}\right) = V\left(\sum_{k\in s_q} d_k(q)\frac{e_{ki}}{m_k}\right) + N^2 o_p(n_N^{-1})$ by Assumption A5, and the result is proven.