University of Granada

Doctorate Program in Mathematics and Statistics

Doctoral Thesis

Inference with data coming from multiple frames: the use of auxiliary information

David Molina Muñoz

Thesis supervised by

Prof. María del Mar Rueda García

Prof. Antonio Arcos Cebrián

Granada, June, 2016

Editor Universidad de Granada. Tesis Doctorales
Autor: David Molina Muñoz
ISBN: 978-84-9125-985-5
URI: http://hdl.handle.net/10481/44077

A mis padres.

El doctorando David Molina Muñoz y los directores de la tesis Dña. María del Mar Rueda García y D. Antonio Arcos Cebrián garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, Junio de 2016.

Directores de la Tesis **Doctorando**

Fdo.: Dña. María del Mar Rueda García

D. Antonio Arcos Cebrián

Fdo.: D. David Molina Muñoz

Contents

List of Figures

List of Tables

Agradecimientos

En la vida, habitualmente, los logros más importantes se consiguen tras recorrer un largo camino. Cuando se alcanza uno de estos grandes hitos conviene girarse sobre uno mismo para poder apreciar el trayecto que has realizado y las personas que te han acompañado en tu caminar. Esta tesis doctoral supone para mí mucho más que un logro: es un sueño cumplido. Por eso, me siento en la obligación de agradecer a todas aquellas personas que, de un modo u otro, han hecho posible su realización.

En primer lugar, me gustaría dar las gracias a mis directores de tesis, los profesores Dña. María del Mar Rueda García y D. Antonio Arcos Cebrián, por aceptar la tutela de este trabajo, por vuestro tiempo, vuestro apoyo y vuestra generosidad. Gracias por compartir conmigo vuestra sabiduría y experiencia como investigadores y por permitirme comprobar de primera mano vuestra excelente calidad humana. También me gustaría agradecer a mis compañeros del Departamento de Estadística e Investigación Operativa y, muy especialmente, a los miembros de la sección departamental del Campus de Cartuja el haberme hecho sentir como en casa durante todo este tiempo.

Grazie anche alla professoressa Maria Giovanna Ranalli e ai miei compagni d'ufficio di Perugia per la loro ospitalità durante il mio indimenticabile soggiorno in Italia.

También tengo que dar las gracias de forma muy cariñosa a mi familia por creer en mí en todo momento (incluso cuando ni yo mismo lo hacía), por apoyarme de forma incondicional y por animarme siempre a seguir adelante. Hacéis que me sienta tremendamente orgulloso y afortunado de formar parte de esta familia. De forma especial, me gustaría agradecer a mis padres, Manuel y Vicenta, la educación que de ellos he recibido, los valores que me han inculcado y el esfuerzo que han realizado para procurarme la mejor de las formaciones.

Y por último, pero no por ello menos importante, gracias a mis amigos por escucharme, por aconsejarme, por aguantarme, por celebrar mis éxitos y sufrir mis fracasos como propios,... En definitiva,

gracias por estar siempre ahí.

Nota: este trabajo de investigación se ha desarrollado al amparo de una beca para la Formación del Profesorado Universitario (FPU) concedida por el Ministerio de Educación, Cultura y Deporte así como de los proyectos de investigación liderados por mis directores de tesis en los cuales he sido partícipe.

Summary

Multiple 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 a multiple frame survey, $Q \geq 2$ sampling frames are available for sampling. Although each of them may be incomplete, it is assumed that, overall, they cover the entire target population. Then, independent samples are selected, one from each frame, under a possibly different sampling design, and information is properly combined to get estimates. Since its emergence, multiple frame sampling theory has experienced a noticeable development and a number of estimators for the total of a continuous variable have been proposed. First proposals were formulated in a dual frame context, i.e. for the case where two frames are available for sampling. Hartley (1962) himself proposed the first dual frame estimator, which was improved by Lund (1968) and Fuller and Burmeister (1972). Bankier (1986) and Kalton and Anderson (1986) and Skinner (1991) proposed dual frame estimators based on new techniques. Skinner and Rao (1996) and Rao and Wu (2010) applied likelihood methods to compute estimators that perform well in complex designs. More recently, Ranalli *et al.* (2015) and Elkasabi *et al.* (2015) used calibration techniques to derive estimators in the dual frame context.

In recent years, a number of works focusing on the estimation in cases with three or more sampling frames has arisen. Lohr and 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 easy to compute. Multiplicity is also used by Rao and Wu (2010) to provide an extension of the pseudo empirical likelihood estimator to the case of more than two frames. In 2011, Singh and Mecatti suggested a class of multiplicity estimators that encompasses all the multiple frames estimators available in the literature by suitably specifying a set of parameters.

However, little attention has been devoted to the study of qualitative variables in a multiple frame context. Qualitative variables are needed to properly represent the responses provided to multiple choice questions, quite frequent in surveys. An important contribution of this thesis is related to the formulation of estimators for the proportions of response variables with discrete outcomes. Estimators for proportions of both multinomial and ordinal response variables have been proposed.

On the other hand, benefits of the multiple frame approach have increased their popularity among the scientific community and now this methodology is frequently used when conducting surveys. Remarkable is the use of dual frame surveys when carrying out telephone surveys. In some subject areas (e.g., electoral), face-to-face surveys have been completely ousted by telephone interviewing. Telephone surveys present some drawbacks with regard to coverage, due to the absence of a telephone in some households and the generalized use of mobile phones, which are sometimes replacing fixed (land) lines entirely. Dual frame telephone surveys that combine Random-Digit-Dialing (RDD) landline telephone samples and cell phone samples are a good solution to that issue since they reduce the noncoverage due to cell-only households in RDD landline telephone surveys. Therefore, in that situations, software for analyzing data coming from dual frame surveys would be very useful. No existing software covered dual frame estimation procedures until *Frames2*, another important contribution of this thesis, was released. *Frames2* is an R package for point and interval estimation in a dual frame context which implements the main estimators for dual frame data proposed so far.

This thesis is presented as a compendium of 4 publications in relation with the contents of the thesis. 3 of the papers are already published in specialized journals and the fourth one is submitted for reviewing. The full version of the papers is included in Appendices A1 - A4, in the second part of the thesis, at the end of it. Previous to the appendix section, a list of chapters that summarize the most important aspects of the papers to facilitate their reading is presented. The first chapter constitutes an introduction to the problem of the estimation in a multiple frame context and a comprehensive overview of the existing approaches for estimating parameters from data coming from a multiple frame survey. Then, the objectives this thesis pursues are enumerated. The methodology used and the most relevant results obtained are presented in Chapters 3 and 4, respectively. Chapter 5 lists the conclusions derived from the results obtained. Finally, Chapter 6 provides some notes on the current research related with the topics addressed in the thesis that is being carried out at present.

Resumen

Las encuestas con marcos múltiples fueron propuestas por Hartley (1962) como un mecanismo para la reducción de los costes de la recolección de datos que conseguía una precisión en los resultados similar a la obtenida con las encuestas de un único marco. En una encuesta con marcos múltiples se dispone de $Q \geq 2$ marcos muestrales. Aunque cada uno de ellos puede ser incompleto, se supone que, conjuntamente, cubren la totalidad de la población de interés. A continuación, se selecciona independientemente una muestra de cada marco considerando diseños muestrales que pueden diferir según el marco y la información recopilada se combina de forma adecuada para obtener estimaciones. Desde su aparición, la teoría de las encuestas con marcos múltiples ha experimentado un importante desarrollo y, como consecuencia, se han formulado numerosos estimadores para el total de una variable continua. Los primeros estimadores se plantearon para el caso en que se dispone de dos marcos para el muestreo (caso conocido como dual frame o de marcos duales). El mismo Hartley (1962) propuso el primer estimador para marcos duales, que fue mejorado posteriormente por Lund (1968) y por Fuller y Burmeister (1972). Bankier (1986) y Kalton y Anderson (1986) y Skinner (1991) sugirieron nuevos estimadores para marcos duales basados en un nuevo enfoque, denominado "single frame". Skinner y Rao (1996) y Rao y Wu (2010) utilizaron técnicas basadas en la verosimilitud para obtener estimadores que se ha demostrado funcionan bien para diseños muestrales complejos. Más recientemente, Ranalli *et al.* (2015) y Elkasabi *et al.* (2015) consideraron métodos de calibración para calcular estimadores para marcos duales.

En los últimos años, han visto la luz un buen número de trabajos de investigación centrados en la estimación en el caso en que se dispone de tres o más marcos muestrales. Lohr y Rao (2006) extendieron algunos de los estimadores propuestos para marcos duales al caso de tres o más marcos muestrales. Mecatti (2007) consideró una nueva metodología basada en la multiplicidad de las unidades de la muestra (es decir, en el número de marcos en los que la unidad se incluye) para proponer un estimador que es muy

sencillo de calcular. Rao and Wu (2010) también consideraron un enfoque basado en la multiplicidad para proponer una extensión al caso de más de dos marcos muestrales del estimador de pseudo verosimilitud empírica que ellos mismos formularon. En 2011, Singh and Mecatti propusieron una clase de estimadores de multiplicidad que englobaba como casos particulares a todos los estimadores para marcos múltiples formulados hasta la fecha. Cada estimador puede obtenerse sin más que ajustar de forma adecuada los valores de un conjunto de parámetros.

Sin embargo, no se ha profundizado demasiado en el estudio de variables cuantitativas en encuestas con marcos múltiples. Este tipo de variables es necesario, por ejemplo, para representar correctamente las respuestas que los individuos muestreados proporcionan a preguntas de respuesta múltiple. Una de las contribuciones más importantes de esta tesis es la formulación de estimadores para la estimación de proporciones de categorías de variables de respuesta discreta. Se ha considerado tanto el caso en que las posibles opciones de la variable respuesta no están ordenadas como aquel otro en que sí existe un determinado orden entre dichas opciones, formulando estimadores adecuados para cada una de las situaciones.

Por otro lado, los beneficios derivados del uso de encuestas con marcos múltiples han hecho que su popularidad se dispare entre la comunidad científica de manera que son muchas las instituciones, tanto públicas como privadas, que se decantan por una metodología basada en marcos múltiples a la hora de llevar a cabo sus encuestas. Especialmente llamativo es el uso de encuestas telefónicas que consideran dos marcos muestrales. En algunas áreas, las encuestas presenciales han sido completamente reemplazadas por las telefónicas. Este el caso, por ejemplo, de las encuestas electorales. Las encuestas telefónicas presentan ciertos inconvenientes relativos a la cobertura, debido a la ausencia de teléfono en algunos hogares y al uso generalizado de teléfonos móviles, los cuales están sustituyendo a los teléfonos fijos en algunos hogares. Una buena solución para este problema viene dada por las encuestas telefónicas con dos marcos muestrales que combinan una muestra de teléfonos fijos y otra de teléfonos móviles obtenidas a través de un marcado automático aleatorio. Mediante esta solución se reduce la falta de cobertura que se obtendría si la encuesta se llevara a cabo únicamente a través de teléfonos fijos producida por aquellos hogares en los que solo se dispone de teléfono móvil. Por todo ello, en este tipo de situaciones, se hace necesario algún software estadístico para el análisis de datos provenientes de encuestas con marcos duales. *Frames2*, otra de las contribuciones más destacadas de esta tesis, es un paquete o librería para el programa estadístico de código abierto R para la estimación puntual y confidencial en encuestas con

marcos duales. Este paquete implementa los principales estimadores para datos provenientes de marcos duales propuestos hasta el momento.

Esta tesis se presenta como un compendio de 4 publicaciones relacionadas con los contenidos de la propia tesis. 3 de los artículos ya se encuentran publicados en revistas especializadas y el cuarto se encuentra sometido a un proceso de revisión. La versión íntegra de los artículos se incluye en los Apéndices A1 - A4, en la segunda parte de la tesis, al final de la misma. Antes de los apéndices, se presentan varios capítulos que resumen los aspectos clave de los artículos para así facilitar la lectura de los mismos. El primer capítulo consituye una introducción al problema de la estimación en encuestas con marcos múltiples así como una revisión de las alternativas existentes para la estimación de parámetros con datos procedentes de este tipo de encuestas. A continuación se enumeran los objetivos que se persiguen con esta tesis. La metodología que se ha seguido y los resultados más importantes que se han obtenido se muestran en los Capítulos 3 y 4, respectivamente. En el Capítulo 5 se listan las conclusiones más relevantes que se derivan de los resultados. Por último, en el Capítulo 6 se exponen brevemente cuáles son los temas que están siendo investigados actualmente en relación con los contenidos de esta tesis.

Part I

Chapter 1

Introduction

Classical sampling theory is based on the existence of a unique sampling frame that includes all the units composing the target population. This is a very strong assumption which is rarely met in practice: populations are constantly changing with new units entering and exiting the population every few time so it is difficult to have an updated list of units from which to draw samples. In these cases, it is said that the sampling frame is incomplete in the sense that it does not include all the units of the population. These differences between the objective population and the sampling frame produce important coverage biases and may affect results due to the non-representativeness of the samples selected from the incomplete frame.

The multiple frame approach has arisen to overcome this issue. The main aim of a multiple frame survey is to estimate the value of a population parameter from the data collected in a sample. To do this, it is assumed that two or more frames are available for sampling and that, overall, they cover the whole target population. So, although each sampling frame can be incomplete treated separately, it is assumed that the union of all of them is complete in the sense that it contains each and every unit of the population. This hypothesis is less restrictive than the one assumed by the classical sampling theory since it is easier to fully reach a population if more than one frame is used. Furthermore, the frames of a multiple frame setting are easier to maintain due to their reduced size in comparison with a single complete frame.

Multiple frame surveys were first introduced by Hartley in 1962 as a device for reducing the costs

derived from data collection while still covering the whole target population. As an example, consider a two frame survey where one of the frames is cheap to sampling from but has an incomplete coverage, whereas the second one is more expensive to sample from but it covers more of the population. In that situation, a multiple frame approach could take advantage of the sampling inexpensiveness of the first frame and of the good coverage of the second to provide better results. Sampling costs depend on many factors as the size of the sample or the mode of interview. In a multiple frame survey these settings can be chosen differently for each frame depending on the peculiarities of each one, so that an appropriate choice may lead to noticeable cost decreases. Other additional technical details, as the sampling design, can also be set independently for each frame.

The use of multiple frames surveys is especially advisable when studying "hidden" or "hard-to-reach" populations. These types of populations are generally named as "rare" populations because individuals composing them present a characteristic which is not frequent in the general population. Although authors have defined the rare populations in several ways, a widely accepted definition for the concept of rare population is the one proposed by Lohr (2009a) who identifies a population as rare when the number of individuals composing it is very small or, even being large, it represents only a small fraction of the global population (usually 10% or less). People suffering from diseases (as AIDS or Alzheimer) or homosexual people are good examples of rare populations. Due to the small representativeness of the individuals of a rare population within the general population, a random sample drawn from this general group will likely include few "rare" elements. In that case, a multiple frame approach may be considered and additional frames containing a high rate of units of the rare population can be sampled in order to increase the sampling size and improve the accuracy of the results. For example, for the population of Alzheimer disease patients, besides sampling in a general sampling frame one could sample in alternative frames as specialized clinics or homes for the elderly to reach a higher number of individuals belonging to the target population. This same reasoning could be applied to elusive or mobile populations which can be seen as a type of rare population due to the difficulties to locate (and therefore, to contact) the individuals composing them. Lots of animal populations (insects, migratory animals, nocturnal behaviour birds,...) are examples of elusive or mobile populations. Finally, a population can be rare both for representing a small part of the global population and for being nomadic as, for example, the homeless people of a particular city.

Multiple frame methodology encompasses all the approaches developed in order to use data coming

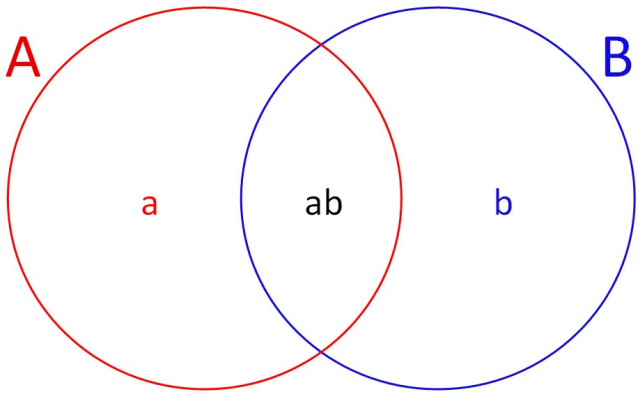


Figure 1.1: Two frames with overlapping

for two or more sampling frames. However, the majority of the literature published so far focuses on the dual frame approach, that is, on the case in which only two frames are available for sampling. The most general situation in a dual frame context is the one depicted in Figure 1.1, in which the two sampling frames present a certain degree of overlapping.

Here, the two frames (generically termed as frame A and frame B) originate three different disjoint non-empty areas or domains: domain a, including population units that belong exclusively to frame A; domain b, including population units that belong only to frame B and, finally, domain ab, including population units that belong simultaneously to both frames. To give an example, let suppose a population of phone users where two sampling frames can be clearly distinguished: the frame A would be, in this case, the one consisting of the users of landline phones and the frame B would be composed of the users of cell phones. We could differentiate, then, the following groups of people: landline-only users, cell-only users and both landline and cell users, which will compose the domains a, b and ab, respectively.

Alternative situations may arise depending on the relative positions of the two frames. Figure 1.2 depicts the case where frame B is totally included in frame A, that is, frame B is a subset of frame A, which is assumed to be complete. In that case, the domain b is an empty set. This would be the case, for example, in a survey where the population of interest is composed of the people facing gambling problems where the frame A is a general population frame and the frame B is composed of the individuals attending therapies to overcome gambling addiction. On the other hand, in figure 1.3 is shown the case in which the two sampling frames coincide, so the only non-empty domain is ab. That kind of situations arise when two different lists of individuals (maybe coming from different sources) of the same target

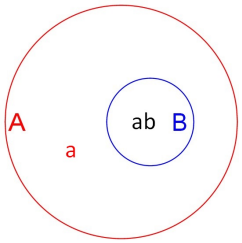


Figure 1.2: Frame B is included in frame A.

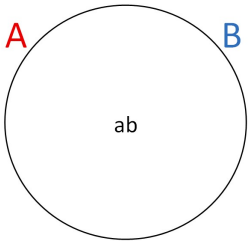


Figure 1.3: Frame A and frame B exactly match.

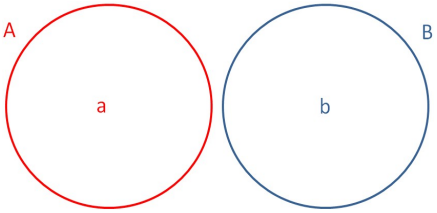


Figure 1.4: Frame A and frame B are disjoint.

population are available from sampling. Finally, the situation in which ab is an empty domain and the two frames do not share units is depicted in figure 1.4. As an example, let suppose that a list of the male individuals and a different one for the female individuals composing the population we are interested in are available. This case can be seen as a special case of a stratified design where each frame represents a stratum. Therefore, it could be analyzed using the customary tools of the stratified sampling and it is not very relevant from a dual frame perspective.

1.1 The estimation problem in dual frame surveys

Without loss of generality, let consider the situation depicted in Figure 1.1, where none of the domains a, b or ab are empty. Let U be a finite population composed of N units labeled from 1 to N, $U = \{1, ..., k, ..., N\}$ and let note the number of units composing frame A and frame B as N_A and N_B, respectively. Similarly, the number of units included in domains a, b and ab are N_a, N_b and N_{ab}, respectively. Let suppose that the parameter of interest is the population total of a continuous variable, and let note that quantity by Y. Therefore $Y = \sum_{k=1}^{N} y_k$, with y_k the value of the variable for the $k-th$ individual of the population. The disjointness of the domains allows us to rewrite the population total as a sum of domain totals, $Y = Y_a + Y_{ab} + Y_b = \sum_{k=1}^{N_a} y_k + \sum_{k=1}^{N_{ab}} y_k + \sum_{k=1}^{N_b} y_k$. To carry out the estimation of the parameter, two random samples are independently drawn, one from each frame. Let denote the sets of units included in the sample drawn from frame A and in the sample drawn from frame B by s_A and s_B, respectively, and let suppose that the number of units selected in each one are n_A and n_B. The final set of units sampled can be computed, then, as $s = s_A \cup s_B$ with size $n = n_A + n_B$. Typically, the sample s_A includes both units from domain a and domain ab and the sample s_B includes units belonging to domain b and to domain ab too, as depicted in the Figure 1.5. Sample s_A can be poststratified as

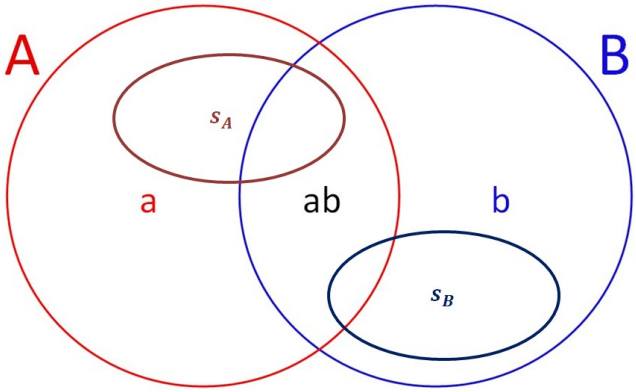


Figure 1.5: Samples drawn from frames.

$s_A = s_a \cup s_{ab}^A$ where s_a and s_{ab}^A are two sets of units of sizes n_a and n_{ab}^A that include the units of s_A belonging to domains a and ab, respectively. Similarly, s_B may be poststratified as $s_B = s_b \cup s_{ab}^B$ with s_b and s_{ab}^B the sets of units of s_B, with sizes n_b and n_{ab}^B, belonging to b and ab.

As indicated before, a different sampling design may be used in each frame, hence probabilities of being selected in the final sample may differ for the population units, depending on the frame each one belongs to. Let note by $\pi_k^A = P(k \in s_A)$ the first order inclusion probability for $k-th$ unit of the frame A, $k = 1, \ldots, N_A$. Therefore, π_k^A indicates the probability of the $k-th$ unit of the frame A to be selected in the sample s_A and, consequently, in the final sample, s. Similarly, $\pi_k^B = P(k \in s_B)$ denotes the first order inclusion probability for the $k-th$ unit of the frame B, $k = 1, \ldots, N_B$. The design weight for each unit is defined as the inverse of its first order inclusion probability, that is, $d_k^A = 1/\pi_k^A$ is the design weight for the $k-th$ unit of the frame A and $d_k^B = 1/\pi_k^B$ is the design weight for the $k-th$ unit of the frame B.

Using the values of the interest variable observed in the units selected in the sample s_A it is possible to compute the customary Horvitz-Thompson estimator of the population total for each of the two domains composing the frame A in the following way:

$$\hat{Y}_a = \sum_{k \in s_A} d_k^A y_k \delta_k(a) = \sum_{k \in s_a} d_k^A y_k \qquad \hat{Y}_{ab}^A = \sum_{k \in s_A} d_k^A y_k \delta_k(ab) = \sum_{k \in s_{ab}^A} d_k^A y_k$$

where $\delta(a)$ and $\delta(ab)$ are the indicator variables for domains a and ab so that $\delta_k(a) = 1$ when the $k-th$ unit of the sample s_A belongs to domain a and 0 otherwise. Equally, $\delta_k(ab) = 1$ whether unit k is included

in domain ab and 0 otherwise.

Similarly, from the information collected in the sample s_B we can compute the Horvitz-Thompson estimator of the population total for domains b and ab as follows:

$$\hat{Y}_b = \sum_{k \in s_B} d_k^B y_k \delta_k(b) = \sum_{k \in s_b} d_k^B y_k \qquad \hat{Y}_{ab}^B = \sum_{k \in s_B} d_k^B y_k \delta_k(ab) = \sum_{k \in s_{ab}^B} d_k^B y_k$$

where, in this case, $\delta_k(b)$ takes the value 1 when the $k-th$ unit of the sample s_B and 0 otherwise. On the other hand, $\delta_k(ab) = 1$ if the $k-th$ unit of s_B belongs to the overlap domain. It is important to note that both $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$ are estimators of the population total in the domain ab but while the first is computed from the information collected in sample s_A, the latter uses the information of s_B.

One could estimate, then, the population total as the sum of the 4 domain estimates, that is $\hat{Y} = \hat{Y}_a + \hat{Y}_{ab}^A + \hat{Y}_{ab}^B + \hat{Y}_b$. The main drawback of this estimator is that it is not an unbiased estimator of Y. Actually,

$$E[\hat{Y}] = E[\hat{Y}_a + \hat{Y}_{ab}^A + \hat{Y}_{ab}^B + \hat{Y}_b] = Y_a + Y_{ab} + Y_{ab} + Y_b = Y + Y_{ab},$$

so this estimator overestimates the real value of the parameter Y. The problem comes from the overlap domain, where two different estimates are considered.

1.2 Different approaches for estimation in dual frame surveys

Units in the overlap domain ab can be selected in both samples s_A and s_B, so the real probability of being included in the final sample s of any unit belonging to the overlap domain is larger than the probability of being included in s_A and also larger than the probability of being included in s_B. In other words, the true first order inclusion probability of the $k-th$ individual of ab is neither π_k^A nor π_k^B but larger. Indeed, the omission of that fact is the cause of the overestimating issue exposed in the previous section. Often, authors refer to this issue as the "multiplicity" or "duplicity issue" due to it is originated by the disregarded appearance of the units in the overlap domain in both sampling frames.

There are several approaches to overcome this problem and to obtain adequate estimates in a dual frame survey. The most used ones are the screening procedure and the dual and single frame approaches.

1.2.1 Screening approach

Screening techniques solve the problem of the overestimation in dual frame surveys just by removing the intersection between frames. That is, they transform scenarios as the depicted in Figures 1.1, 1.2 and 1.3 in a scenario similar to the shown Fn figure 1.4. To do this, they remove the units in the overlap domain from one of the sampling frames removing, then, the overlap as well. As mentioned previously, at this point, stratified estimators can be considered for estimating the parameter of interest from data collected in the samples coming from each frame, which represents a stratum in this case.

The main drawback of this approach is the necessity of identifying the population units that are included simultaneously in both sampling frames to properly remove them from one of the frames. In most cases it is impossible to know beforehand the domain each population unit belongs to. Usually, that information is only available for the sample units, which are assigned to the corresponding domain once the needed information is collected during the interview. However, there are situations where the screening process is feasible. González-Villalobos and Wallace (1996) presented an example of a screening survey where both a land area frame and a list frame of farms were considered.

Benefits of the screening process have been questioned in literature. Indeed, González-Villalobos and Wallace (1996) themselves refer to the screening as "an operation that requires special attention and resources". On the other hand, Mecatti (2014) affirmed that screening operations can be resource-consuming, error-prone, and essentially amount to missed opportunity to collect data from a willing participant. Kennedy (2007) discussed the effects of screening in a dual frame phone survey, finding that the screening techniques may increase the nonresponse error leading to different results depending on the frame the repeated units are removed from.

Whatever the case, since any screening survey can be seen as a particular case of a stratified survey, we are not discussing the topic in depth.

1.2.2 Dual frame approach

The dual frame approach suggests a convex combination of the two overlap estimates to obtain an unbiased global estimator of the parameter of interest. That is, the population total in the common domain, Y_{ab}, is estimated as $\hat{Y}_{ab} = \theta\hat{Y}^A_{ab} + (1-\theta)\hat{Y}^B_{ab}$, where the weighting parameter θ is between 0 and 1, $\theta \in [0,1]$. When the value 0 is selected, $\hat{Y}_{ab} = \hat{Y}^B_{ab}$ so only the information regarding units of

the overlap domain that has been collected from the units selected in the sample drawn from frame B is used to estimate the population total in domain ab. Similarly, when $\theta = 1$, $\hat{Y}_{ab} = \hat{Y}^A_{ab}$. Both situations can be considered as particular cases of screening and, therefore, of stratification. For this reason, from now on we are considering only cases where $\theta \in (0, 1)$. Summarizing, the dual frame approach weights the two estimates of the parameter of interest of the overlap domain to avoid overestimation issues. The population total is estimated, then, through the estimator

$$\hat{Y} = \hat{Y}_a + \theta\hat{Y}^A_{ab} + (1-\theta)\hat{Y}^B_{ab} + \hat{Y}_b, \tag{1.1}$$

which is now clearly unbiased for Y.

Estimator (1.1) can be rewritten as $\hat{Y} = \sum_{k\in s} d^\circ_k y_k$, where the weights d°_k are defined for the units in the poststratified samples as follows:

$$d^\circ_k = \begin{cases} d^A_k & \text{if } k \in s_a \\ \theta d^A_k & \text{if } k \in s^A_{ab} \\ (1-\theta)d^B_k & \text{if } k \in s^B_{ab} \\ d^B_k & \text{if } k \in s_b \end{cases} \tag{1.2}$$

Unlike the screening, the dual frame approach can always be applied since it does not require any previous or additional information but only the choice of the parameter θ. Along the years, authors have proposed different procedures for selecting the value of θ, yielding to different estimators as it will be shown in the subsequent section. These techniques encompass simple options that select a fixed value for θ and more complex approaches where θ is determined to optimize, in some sense, the estimates.

1.2.3 Single frame approach

The idea underlying the single frame approach (Bankier (1986) and Kalton and Anderson (1986)) is to adjust the inclusion probabilities (or, equivalently, the design weights) of the units of the overlap domain to properly take into account the fact that they may be selected both in samples s_A and s_B. After that, the units composing the two frames may be combined into a single dataset (hence the name of the technique).

As mentioned before, a population unit belonging to the domain ab has the chance of being included

in the sample s_A and in the sample s_B too. So, the real first order inclusion probability for the $k-th$ unit of the overlap domain is $\pi_k^A + \pi_k^B$. Then, new sets of adjusted weights can be defined for the units of the poststratified samples as follows

$$\tilde{d}_k^A = \begin{cases} d_k^A & \text{if } k \in s_a \\ (1/d_k^A + 1/d_k^B)^{-1} & \text{if } k \in s_{ab}^A \end{cases} \qquad \tilde{d}_k^B = \begin{cases} d_k^B & \text{if } k \in s_b \\ (1/d_k^A + 1/d_k^B)^{-1} & \text{if } k \in s_{ab}^B \end{cases}$$

or, summarizing,

$$\tilde{d}_k = \begin{cases} d_k^A & \text{if } k \in s_a \\ (1/d_k^A + 1/d_k^B)^{-1} & \text{if } k \in s_{ab}^A \cup s_{ab}^B \\ d_k^B & \text{if } k \in s_b \end{cases} \tag{1.3}$$

The main problem with the computation of the weights in (1.3) lies in the requirement of the knowledge of the first order inclusion probabilities for the units composing the common domain both under the sampling design used in frame A and the sampling design used in frame B, which is not always the case, especially when complex sampling designs are considered.

1.3 Existing estimators in dual frame surveys

Since Hartley presented the dual frame methodology in 1962, a number of estimators have been formulated to estimate parameters using data coming from two frames, both under dual and single frame approaches.

The simplest estimator is computed by selecting a predetermined value between 0 and 1 for the weighting parameter θ and then by substituting it in the expression (1.1). The resulting estimator is often called "fixed weight estimator". Different criteria can be considered for the choice of θ based on previous studies or on known information about the behavior of the interest variable in the overlap domain. A value of θ which usually provide good results (Brick *et al.* (2006)) is $\theta = 1/2$ and the corresponding estimator in that case can be written as

$$\hat{Y}_{FW} = \hat{Y}_a + \frac{1}{2}\hat{Y}_{ab}^A + \frac{1}{2}\hat{Y}_{ab}^B + \hat{Y}_b = \hat{Y}_a + \frac{1}{2}(\hat{Y}_{ab}^A + \hat{Y}_{ab}^B) + \hat{Y}_b \tag{1.4}$$

Hartley (1962) generalized the idea of the fixed weight estimator, proposing a class of estimators for

a simple random sampling design in each frame in the form

$$\hat{Y}_H = \hat{Y}_a + \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b \tag{1.5}$$

with $\theta \in (0,1)$. Hartley himself computed in 1974 the optimal value of θ in the sense that the asymptotic variance of the estimator is minimized. That optimal value can be written as follows

$$\theta_H = \frac{V(\hat{Y}_{ab}^B) + Cov(\hat{Y}_b, \hat{Y}_{ab}^B) - Cov(\hat{Y}_a, \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)}$$

where $V()$ and $Cov()$ represent the population variance of an estimator and the population covariance between two estimators, respectively. Usually, these population variances and covariances are unknown and have to be estimated from the information available in samples s_A and s_B, resulting in the following estimator of θ_H:

$$\hat{\theta}_H = \frac{\hat{V}(\hat{Y}_{ab}^B) + \widehat{Cov}(\hat{Y}_b, \hat{Y}_{ab}^B) - \widehat{Cov}(\hat{Y}_a, \hat{Y}_{ab}^A)}{\hat{V}(\hat{Y}_{ab}^A) + \hat{V}(\hat{Y}_{ab}^B)}$$

Since $\hat{\theta}_H$ is consistent for θ_H, Hartley estimator is asymptotically optimal among all estimators of the form $\hat{Y}_a + \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b$. However, this estimator presents some important drawbacks. First, the fact that θ_H depends on values of the main variable makes the estimator internally inconsistent in the sense that the sum of the estimates in the subsets conforming a partition of the population does not coincide with the estimation for the entire population. For example, let suppose that $\hat{Y}_1$ estimates the men in the population, $\hat{Y}_2$ estimates the women in the population and $\hat{Y}_3$ estimates the total number of people in the population. Then, it would be desirable that $\hat{Y}_1 + \hat{Y}_2 = \hat{Y}_3$. But this is only true when the estimates are computed using an internally consistent estimator, which is not the case of the Hartley estimator. Furthermore, as indicated by Lohr (2009b), when the absolute value of $\widehat{Cov}(\hat{Y}_b, \hat{Y}_{ab}^B)$ or $\widehat{Cov}(\hat{Y}_a, \hat{Y}_{ab}^A)$ is large, the values of $\hat{\theta}_H$ may fall outside the interval $(0,1)$.

Lund (1968) improved Hartley estimator by considering the random division of the frame sample sizes among domains actually achieved in the sample. The estimator he proposed is given by

$$\hat{Y}_L = \hat{Y}_a + \left(\theta\frac{N_A}{n_A}n_{ab}^A + (1-\theta)\frac{N_B}{n_B}n_{ab}^B\right)\frac{\hat{Y}_{ab}}{n_{ab}} + \hat{Y}_b, \tag{1.6}$$

where n_{ab}^A and n_{ab}^B represent the number of units of the samples s_A and s_B, respectively, belonging to

domain ab. On the other hand, $\hat{Y}_{ab} = \sum_{k \in s^A_{ab}} y_k + \sum_{k \in s^B_{ab}} y_k$ and $n_{ab} = n^A_{ab} + n^B_{ab}$. Again, θ is a constant in the interval $(0, 1)$. As it was the case with the Hartley estimator, computation of the optimum value of θ that minimizes the asymptotic variance of the Lund estimator involves unknown population quantities, so that it has to be estimated from the information collected in the sample. Lund suggested the estimator

$$\hat{\theta}_L = \left(\left(\frac{N_A n_a}{n_A^2} + \frac{N_B n_b}{n_B^2} \right) \frac{y_{ab}}{n_{ab}} \right)^{-1} \left(\frac{\hat{Y}_a}{n_A} + \frac{N_B n_b}{n_B^2} \frac{\hat{Y}_{ab}}{n_{ab}} - \frac{\hat{Y}_b}{n_B} \right),$$

with n_a and n_b the number of sample units belonging to domain a and to domain b, respectively. $\hat{\theta}_L$ depends on the values of the main variable, making the estimator internally inconsistent. It can be proved that the variance of the Lund estimator is always less or equal than the variance of the Hartley estimator, irrespective of the value of θ.

Fuller and Burmeister (1972) introduced information about the estimation of the unknown overlap domain size, N_{ab}, to further improve the Hartley estimator. The estimator they proposed can be written as

$$\hat{Y}_{FB} = \hat{Y}_a + \beta_1 \hat{Y}^A_{ab} + (1 - \beta_1) \hat{Y}^B_{ab} + \hat{Y}_b + \beta_2 (\hat{N}^A_{ab} - \hat{N}^B_{ab}), \tag{1.7}$$

where $\hat{N}^A_{ab} = \sum_{k \in s^A_{ab}} d^A_k$ and $\hat{N}^B_{ab} = \sum_{k \in s^B_{ab}} d^B_k$ are the estimates of the overlap domain size computed from the information collected in s_A and in s_B, respectively. The authors also shown that

$$\begin{bmatrix} \beta_{1_{FB}} \\ \beta_{2_{FB}} \end{bmatrix} = - \begin{bmatrix} V(\hat{Y}^A_{ab} - \hat{Y}^B_{ab}) & Cov(\hat{Y}^A_{ab} - \hat{Y}^B_{ab}, \hat{N}^A_{ab} - \hat{N}^B_{ab}) \\ Cov(\hat{Y}^A_{ab} - \hat{Y}^B_{ab}, \hat{N}^A_{ab} - \hat{N}^B_{ab}) & V(\hat{N}^A_{ab} - \hat{N}^B_{ab}) \end{bmatrix}^{-1} \times \begin{bmatrix} Cov(\hat{Y}_a + \hat{Y}_b + \hat{Y}^B_{ab}, \hat{Y}^A_{ab} - \hat{Y}^B_{ab}) \\ Cov(\hat{Y}_a + \hat{Y}_b + \hat{Y}^B_{ab}, \hat{N}^A_{ab} - \hat{N}^B_{ab}) \end{bmatrix}$$

are the optimal values for β_1 and β_2 in the sense of minimization the variance of the estimator. In practice, values $\beta_{1_{FB}}$ and $\beta_{2_{FB}}$ are generally unknown and have to be estimated from sample data, resulting in different values depending on the response variable. Therefore, this estimator is also internally inconsistent.

The Hartley and the Lund estimators can be seen as particular cases of the Fuller and Burmeister estimator, presenting the latter the smallest asymptotic variance of all.

The estimators described so far were proposed following a dual frame approach. Bankier (1986) and

Kalton and Anderson (1986) introduced the single frame methodology as an alternative to the dual frame approach. They suggested to group the units of the two samples in a single dataset and use the modified weights defined in (1.3) to propose the estimator

$$\hat{Y}_{SF} = \sum_{k \in s_A} y_k \tilde{d}_k + \sum_{k \in s_B} y_k \tilde{d}_k = \sum_{k \in s} y_k \tilde{d}_k \tag{1.8}$$

This estimator is easy to compute and has the advantage of using the same set of weights regardless of the main variable considered. On the other hand, it presents the inconvenient of requiring the knowledge of the inclusion probabilities of the units belonging to the intersection domain under the sampling designs considered in both frames and not only under the one used in the frame the unit has been sampled from. This may be a challenging deal when the samples selected are not self-weighted.

The Bankier-Kalton-Anderson estimator (often called single frame estimator) can be improved when the frame sizes N_A and N_B are known. In this case, several procedures can be used to incorporate that auxiliary information to the estimation process. Bankier (1986) calibrated the single frame estimator to the frames sizes using an iterative algorithm based on the raking ratio estimation. Rao and Skinner (1996) proved that the raking procedure converges and provided the explicit form of the estimator

$$\hat{Y}_{SFRR} = \frac{N_A - \hat{N}_{ab}^{rake}}{\hat{N}_a} \hat{Y}_a + \frac{N_B - \hat{N}_{ab}^{rake}}{\hat{N}_b} \hat{Y}_b + \frac{\hat{N}_{ab}^{rake}}{\hat{N}_{abS}} \hat{Y}_{abS}, \tag{1.9}$$

where $\hat{Y}_{abS} = \sum_{k \in s_{ab}^A} y_k \tilde{d}_k + \sum_{k \in s_{ab}^B} y_k \tilde{d}_k$, $\hat{N}_{abS} = \sum_{k \in s_{ab}^A} \tilde{d}_k + \sum_{k \in s_{ab}^B} \tilde{d}_k$, $\hat{N}_a = \sum_{k \in s_a} d_k^A, \hat{N}_b = \sum_{k \in s_b} d_k^B$ and $\hat{N}_{ab}^{rake}$ is the smallest root of the quadratic equation $\hat{N}_{abS} x^2 - (\hat{N}_{abS}(N_A + N_B) + \hat{N}_{aS}\hat{N}_{bS})x + \hat{N}_{abS} N_A N_B = 0$, with $\hat{N}_{aS} = \sum_{k \in s_a} \tilde{d}_k$ and $\hat{N}_{bS} = \sum_{k \in s_b} \tilde{d}_k$.

Alternatively, regression estimation can be considered for adjusting the frame sizes N_A and N_B. Lohr and Rao (2000) proposed the following estimator

$$\hat{Y}_{SFReg} = \hat{Y}_{SF} + \hat{\beta}_S^{'}(N_A - \hat{N}_{AS}, N_B - \hat{N}_{BS})^{'}, \tag{1.10}$$

where $\hat{\beta}_S^{'} = -\widehat{Cov}(\hat{N}_{AS}/\hat{V}(\hat{N}_{AS}), \hat{N}_{BS}/\hat{V}(\hat{N}_{BS}), \hat{Y}_{SF})$, with $\hat{N}_{AS} = \sum_{k \in s_A} \tilde{d}_k$ and $\hat{N}_{BS} = \sum_{k \in s_B} \tilde{d}_k$. On the other hand, $\hat{V}(\hat{N}_{AS})$ and $\hat{V}(\hat{N}_{BS})$ are the estimated variances for the estimators $\hat{N}_{AS}$ and $\hat{N}_{BS}$, respectively.

Skinner and Rao (1996) used a pseudo maximum likelihood approach to extend the Fuller and Burmeis-

ter estimator (1.7), which assumes a simple random sampling in each frame, to the case of complex sampling designs. Their results were based on the paper of Skinner (1991), who shown that the Fuller and Burmeister estimator can be derived following maximum likelihood principles. As a result, they proposed the following estimator

$$\begin{aligned}\hat{Y}_{PML} = &\frac{N_A - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a}\hat{Y}_a + \frac{N_B - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b}\hat{Y}_b \\ &+ \frac{\hat{N}_{ab}^{PML}(\theta)}{\theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B}[\theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B],\end{aligned} \tag{1.11}$$

where $\hat{N}_{ab}^{PML}(\theta)$ is the smallest of the roots of quadratic equation $[\theta/N_B + (1-\theta)/N_A]x^2 - [1 + \theta\hat{N}_{ab}^A/N_B + (1-\theta)\hat{N}_{ab}^B/N_A]x + \theta\hat{N}_{ab}^A + (1-\theta)\hat{N}_{ab}^B = 0$ and $\theta \in (0,1)$. It is also shown that the following value for θ

$$\theta_{PML} = \frac{N_a N_B V(\hat{N}_{ab}^B)}{N_a N_B V(\hat{N}_{ab}^B) + N_b N_A V(\hat{N}_{ab}^A)}$$

minimizes the variance of $\hat{Y}_{PML}$. Values N_a, N_b and the variances involved in the computation of θ_{PML} are usually unknown and must be estimated from the sample data yielding to the estimated value $\hat{\theta}_{PML}$ which is substituted in (1.11). The authors suggested the following approximation for θ_{PML} based only on the variances of the estimators of the overlap domain:

$$\phi_{PML} = \frac{V(\hat{N}_{ab}^B)}{V(\hat{N}_{ab}^B) + V(\hat{N}_{ab}^A)},$$

which is easier to compute. In order to calculate the parameter θ_{PML} it is required all the three domains a, b or ab to be nonempty and variances $V(\hat{N}_{ab}^A)$ and $V(\hat{N}_{ab}^B)$ to be positive. Otherwise, serious difficulties may arise and alternative approaches have to be considered. This is the situation, for example, when the sampling frames are placed as shown in Figure 1.2 or in Figure 1.3. In such cases, Lohr and Rao (2006) proposed calculate the value of θ using average design effects for a fixed subset of important variables.

θ_{PML} does not depends on values of the response variable which assures the internal consistency of the estimator. On the other hand, although the pseudo maximum likelihood may not be optimal under complex sampling designs, Skinner and Rao (1996) and Lohr and Rao (2006) found that it has small mean squared error and works well in many situations.

Usually, additional information about auxiliary variables is collected when conducting surveys. This

information may be taken into account in the estimation process resulting in a considerable improvement of the accuracy of the estimates when there is a significant relationship between the set of auxiliary and the main variables. In a dual frame context, the most general situation is when a totally different set of auxiliary variables is considered in each frame. In that case, one may note as $\check{\mathbf{X}}^A = (\mathbf{X}_1^A, ..., \mathbf{X}_p^A)$ the set of the p variables composing the auxiliary information in frame A, so that the vector $\mathbf{x}_k^A = (x_{k_1}^A, ..., x_{k_p}^A)$ contains the values of the variables $\check{\mathbf{X}}^A$ for the $k-th$ member of the frame A. Similarly, in frame B a set of r additional variables is considered and we note it as $\check{\mathbf{X}}^B = (\mathbf{X}_1^B, ..., \mathbf{X}_r^B)$, being $\mathbf{x}_l^B = (x_{l_1}^B, ..., x_{l_r}^B)$ the combination of values of $\check{\mathbf{X}}^B$ for the $l-th$ individual belonging to frame B. When sampling, together with the main variable, the corresponding auxiliary variables are observed. So, for the $k-th$ individual interviewed in s_A, values $(y_k, \mathbf{x}_{k_1}^A, ..., \mathbf{x}_{k_p}^A)$ are collected and, analogously values of $(y_l, \mathbf{x}_{l_1}^B, ..., \mathbf{x}_{l_r}^B)$ are noted for the l-th individual of s_B.

In 2010, Rao and Wu formulated a pseudo empirical likelihood estimator for the mean of a quantitative variable which is able to deal with auxiliary information. The estimator is in the form

$$\hat{\bar{Y}}_{PEL} = \frac{N_a}{N}\hat{\bar{Y}}_a + \theta\frac{N_{ab}}{N}\hat{\bar{Y}}_{ab}^A + (1-\theta)\frac{N_{ab}}{N}\hat{\bar{Y}}_{ab}^B + \frac{N_b}{N}\hat{\bar{Y}}_b, \tag{1.12}$$

where, in this case, $\hat{\bar{Y}}_a = \sum_{k\in s_a}\hat{p}_{ak}y_k, \hat{\bar{Y}}_{ab}^A = \sum_{k\in s_{ab}^A}\hat{p}_{abk}^A y_k, \hat{\bar{Y}}_{ab}^B = \sum_{k\in s_{ab}^B}\hat{p}_{abk}^B y_k$ and $\hat{\bar{Y}}_b = \sum_{k\in s_b}\hat{p}_{bk}y_k$. The four sets of probability measures $\mathbf{p}_a = (\hat{p}_{a1}, ..., \hat{p}_{an_a})'$, $\mathbf{p}_{ab}^A = (\hat{p}_{ab1}^A, ..., \hat{p}_{abn_{ab}^A}^A)'$, $\mathbf{p}_{ab}^B = (\hat{p}_{ab1}^B, ..., \hat{p}_{abn_{ab}^B}^B)'$ and $\mathbf{p}_b = (\hat{p}_{b1}, ..., \hat{p}_{bn_b})'$ are such that maximize the following pseudo empirical likelihood function

$$\begin{aligned} l(\mathbf{p}_a, \mathbf{p}_{ab}^A, \mathbf{p}_{ab}^B, \mathbf{p}_b) = n\Bigg(&\frac{N_a}{N}\sum_{k\in s_a}\tilde{d}_{ak}log(p_{ak}) + \theta\frac{N_{ab}}{N}\sum_{k\in s_{ab}^A}\tilde{d}_{abk}^A log(p_{abk}^A) \\ &+(1-\theta)\frac{N_{ab}}{N}\sum_{k\in s_{ab}^B}\tilde{d}_{abk}^B log(p_{abk}^B) + \frac{N_b}{N}\sum_{k\in s_b}\tilde{d}_{bk}log(p_{bk})\Bigg) \end{aligned}$$

subject to the constraints

$$\sum_{k\in s_a}p_{ak} = \sum_{k\in s_{ab}^A}p_{abk}^A = \sum_{k\in s_{ab}^B}p_{abk}^B = \sum_{k\in s_b}p_{bk} = 1. \tag{1.13}$$

The weights $\tilde{d}_{ak} = d_k^A / \sum_{k\in s_a} d_k^A, \tilde{d}_{abk}^A = d_k^A / \sum_{k\in s_{ab}^A} d_k^A, \tilde{d}_{abk}^B = d_k^B / \sum_{k\in s_{ab}^B} d_k^B$ and $\tilde{d}_{bk} = d_k^B / \sum_{k\in s_b} d_k^B$ are the normalized weights by domains. Again, θ is a weighting parameter between 0 and 1.

The authors also impose the additional constraint

$$\sum_{k \in s_{ab}^{A}} p_{abk}^{A} y_k = \sum_{k \in s_{ab}^{B}} p_{abk}^{B} y_k \tag{1.14}$$

to make sure that the two estimates for the mean of the variable in the common domain coincide, granting consistency to the estimator. They found that the optimal value of θ that minimizes the asymptotic variance of the estimator is given by

$$\theta_{PEL} = \frac{V(\hat{\bar{Y}}_{ab}^{B})}{V(\hat{\bar{Y}}_{ab}^{A}) + V(\hat{\bar{Y}}_{ab}^{B})},$$

which generally depends on the values of the main variable leading to internal inconsistencies. As an alternative, they propose the use of (1.3) as weighting parameter.

Note that to compute the estimator (1.12) it is assumed that frame sizes, N_A and N_B, and overlap domain size, N_{ab}, are known. However, in their paper, Rao and Wu indicate how to estimate these values when one or several of then are not known.

The pseudo empirical likelihood estimator can incorporate auxiliary population information into inference through additional constraints. So, when the vector $\mathbf{X}^A$ of population totals of the variables $\check{\mathbf{X}}^A$ is known, the constraint

$$\frac{N_a}{N} \sum_{k \in s_a} p_{ak} \mathbf{x}_k^A + \theta \frac{N_{ab}}{N} \sum_{k \in s_{ab}^{A}} p_{abk}^{A} \mathbf{x}_k^A = \frac{\mathbf{X}^A}{N}$$

is considered together with constraints (1.13) and (1.14) when maximizing the pseudo empirical likelihood function. A similar constraint can be posed in the case where $\mathbf{X}^B$, the vector of population totals of $\check{\mathbf{X}}^B$, is known.

Recently, Ranalli *et al.* (2015) extended the calibration techniques originally proposed by Deville and Särndal (1992) for an only frame to the case of two sampling frames. As a result, they suggested two different model calibrated estimators: one constructed under the dual frame approach and another one formulated following the single frame methodology.

Under the assumption of frame sizes, N_A and N_B, and overlap domain size, N_{ab}, known, the dual frame calibration estimator can be written as

$$\hat{Y}_{CAL_{DF}} = \sum_{k \in s_A} w_k^{\circ} y_k + \sum_{k \in s_B} w_k^{\circ} y_k = \sum_{k \in s} w_k^{\circ} y_k, \tag{1.15}$$

where the weights w_k° are such that minimize $\sum_{k \in s} G(w_k^{\circ}, d_k^{\circ})$, with $G(\cdot, \cdot)$ a particular distance measure, subject to

$$\sum_{k \in s_a} w_k^{\circ} = N_a \qquad \sum_{k \in s_{ab}^A} w_k^{\circ} = \theta N_{ab} \qquad \sum_{k \in s_{ab}^B} w_k^{\circ} = (1-\theta) N_{ab} \qquad \sum_{k \in s_b} w_k^{\circ} = N_b \tag{1.16}$$

being $\theta \in (0, 1)$ fixed. The authors suggest the choice of values for θ that do not depend on the values of the main variable, as is the case of (1.3).

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

As with the estimator (1.12), the dual frame calibration estimator is able to incorporate information about auxiliary variables to the estimation process. Supposing that $\mathbf{X}^A$, the vector of population totals for the set of variables $\check{\mathbf{X}}^A$ observed for the units of frame A, is known, one should consider, in addition to (1.16), the calibration constraint

$$\sum_{k \in s_a} w_k^{\circ} \mathbf{x}_k^A + \sum_{k \in s_{ab}^A} w_k^{\circ} \mathbf{x}_k^A + \sum_{k \in s_{ab}^B} w_k^{\circ} \mathbf{x}_k^A = \mathbf{X}^A. \tag{1.17}$$

Note that formulation of (1.17) requires the knowledge of values of $\mathbf{x}_k^A$ for the units of s_{ab}^B. Although these units are included in the sample drawn from frame B they also form part of the frame A (indeed, they belong to both frames, since they are located in the overlap domain ab). A constraint similar to (1.17) is formulated when the population totals about auxiliary variables of the frame B are available.

In the context of N_A, N_B and N_{ab} known, the second calibration estimator developed by Ranalli *et. al* (2015), often referred as single frame calibration estimator, is given by the following expression:

$$\hat{Y}_{CAL_{SF}} = \sum_{k \in s_A} \tilde{w}_k y_k + \sum_{k \in s_B} \tilde{w}_k y_k = \sum_{k \in s} \tilde{w}_k y_k, \tag{1.18}$$

where, in this case, weights $\tilde{w}_k$ are such that minimize $G(\tilde{w}_k, \tilde{d}_k)$, being again $G(\cdot, \cdot)$ a particular distance

measure, subject to the following constraints:

$$\sum_{k\in s_a} \tilde{w}_k = N_a \qquad \sum_{k\in s_{ab}^A} \tilde{w}_k + \sum_{k\in s_{ab}^B} \tilde{w}_k = N_{ab} \qquad \sum_{k\in s_b} \tilde{w}_k = N_b \tag{1.19}$$

If the vector of population totals $\mathbf{X}^A$ of the auxiliary variables $\check{\mathbf{X}}^A$ observed in the frame A is known, then a constraint similar to (1.17) is also considered, but replacing w_k° by $\tilde{w}_k$. The same comment is applicable if the vector $\mathbf{X}^B$ is known.

Both estimators present quite a few similarities, mainly due to they have been constructed following a similar procedure. However, they also show some differences. The most noticeable one lies in the use of weights d_k° as starting weights for the calibration used by the dual frame estimator instead of the weights $\tilde{d}_k$ used by the single frame one.

The two calibration estimators have been defined assuming the knowledge of the frame sizes and the overlap domain size, which can be quite restrictive in some cases. Indeed, unlike the frame sizes, which are usually known when conducting a dual frame survey, common domain size is not always available. The authors also indicate the modifications that must be carried out in the set of constraints (1.16) and (1.19) to encompass this situation. They also noticed that some of the estimators exposed so far, as (1.9) or (1.12), can be seen as special cases of calibration estimators considering appropriate combinations of distance measures and sets of constraints.

Elkasabi *et al.* (2015) also used a calibration approach to formulate the so called joint calibration estimator, which may be expressed as

$$\hat{Y}_{JCE} = \sum_{k\in s_A} w_k^\star y_k + \sum_{k\in s_B} w_k^\star y_k = \sum_{k\in s} w_k^\star y_k, \tag{1.20}$$

with weights $w_k^\star$ minimizing $G(w_k^\star, d_k)$, a specific distance measure with respect to original design weights subject to the constraints

$$\sum_{k\in s} w_k^\star = N \qquad \sum_{k\in s_a} w_k^\star = N_a \qquad \sum_{k\in s_b} w_k^\star = N_b \tag{1.21}$$

Again, the population size N and the sizes of domains a and b are supposed to be known. This is equivalent to know the frame sizes N_A and N_B and the overlap domain size N_{ab}. The authors provide

the explicit form for the weights $w_k^\star$ when the linear distance is considered.

If, additionally, the vector $\mathbf{X}^A$ of population totals of the variables $\check{\mathbf{X}}^A$ is known, an extra constraint, similar to (1.17) where weights w_k° are replaced by $w_k^\star$, is considered together with (1.21) when searching the new set of weights. The same argument applies when $\mathbf{X}^B$ is known.

The joint calibration estimator is asymptotically design unbiased conditional on the strong relationship between the estimation variable and the auxiliary variables employed in the calibration.

1.4 Variance estimation in dual frame surveys

The estimation of the variance of estimators presented in the previous section is not always straightforward. For the majority of the estimators internally consistent, the estimation of the variance may be carried out from the independence of the samples s_A and s_B. This is the case of estimators (1.1)(with a fixed value of θ) or (1.8). In these situations, the estimated variance can be obtained as the sum of the estimated variances of the estimators for the two samples. Thus, estimated variance of estimator (1.1) may be expressed as

$$\hat{V}(\hat{Y}) = \hat{V}(\hat{Y}_a + \theta\hat{Y}_{ab}^A) + \hat{V}((1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b) \tag{1.22}$$

The two estimated variances composing the sum may be obtained using the sampling design, so $\hat{V}(\hat{Y})$ can be easily computed. Likewise, variance of (1.8) can be written as

$$\hat{V}(\hat{Y}_{SF}) = \hat{V}\left(\sum_{k\in s_A} y_k\tilde{d}_k\right) + \hat{V}\left(\sum_{k\in s_B} y_k\tilde{d}_k\right) \tag{1.23}$$

Nevertheless, this reasoning cannot be followed to estimate the variance of the pseudo maximum likelihood estimator (1.11) and of the internally inconsistent estimators ((1.5), (1.6), (1.7) and (1.10)). The computation of all these estimators involves the calculation of a weighting parameter that depends directly on the values of the study variable or on estimated variances or covariances from the frames. This generates an additional variability that must be captured when estimating the variance. A similar comment is applied to the calibration estimators (1.15), (1.18) and (1.20) (and, therefore, for (1.9) and (1.12)), since an extra variability that should be taken into account in the variance estimation is produced when calibrating weights to population quantities. Each author addresses this issue by suggesting a specific variance estimator for the estimator they propose, which leads to difficulties when comparing

estimators.

In that cases, alternative techniques as Taylor linearization, jackknife or bootstrap have been proposed to estimate the variance of the estimators in a unified way. Skinner and Rao (1996) used a method based on the Taylor linearization to estimate the variance of the estimator (1.11). Lohr and Rao (2000) discussed that procedure in the more general situation in which the parameter can be written as a function, let say g, of the population means in the two frames. In that context, the linearization variance estimator of a generic estimated parameter, $\hat{\eta}$, is defined as

$$\hat{V}_{Lin}(\hat{\eta}) = g^{A'} S^A g^A + g^{B'} S^B g^B, \tag{1.24}$$

being g^A and g^B the vectors of first partial derivatives of g in frame A and B, respectively. On the other hand, S^A and S^B are the estimated covariance matrices of the population totals estimated from frame A and B.

It is shown that, under certain regularity conditions, the linearization variance estimator is consistent but it presents the important drawback that derivatives should be calculated separately for each different parameter.

Alternatively, one can consider jackknife techniques, originally proposed by Quenouille (1949, 1956) (see Wolter (2007) for a detailed description of this method in survey sampling) and extended to dual frame surveys by Lohr and Rao (2000), which can be used to estimate variances irrespective of the type of estimator allowing us to compare estimated efficiency for different estimators.

For a non stratified design in each frame, the jackknife estimator of the variance for any of the estimators described, generically denoted by $\hat{\eta}$, is given by

$$\hat{V}_{Jack}(\hat{\eta}) = \frac{n_A - 1}{n_A} \sum_{i \in s_A} (\hat{\eta}^A(i) - \overline{\eta}^A)^2 + \frac{n_B - 1}{n_B} \sum_{j \in s_B} (\hat{\eta}^B(j) - \overline{\eta}_c^B)^2, \tag{1.25}$$

with $\hat{\eta}^A(i)$ the value of estimator $\hat{\eta}$ after dropping unit i from s_A and $\overline{\eta}^A$ the mean of values $\hat{\eta}^A(i)$. Similarly, one can define $\hat{\eta}^B(j)$ and $\overline{\eta}^B$.

Jackknife may present an important bias when designs are without replacement. One could, then, incorporate an approximate finite-population correction to estimation to achieve unbiasedness. For example, assuming that a finite-population correction is needed in frame A, a modified jackknife estimator of

variance, $\hat{V}^{\star}_{Jack}(\hat{\eta})$, can be calculated by replacing $\hat{\eta}^A(i)$ in (1.25) with $\hat{\eta}^{A*}(i) = \hat{\eta}_c + \sqrt{1-\overline{\pi}_A}(\hat{\eta}^A(i) - \hat{\eta})$, where $\overline{\pi}_A = \sum_{k \in s_A} \pi_k^A / n_A$.

Consider now a stratified design in each frame, where frame A is divided into H strata and frame B is divided into L strata. From stratum h of frame A, a sample of n_{Ah} units from the N_{Ah} population units in the stratum is drawn. Similarly, in stratum l of frame B, one selects n_{Bl} units from the N_{Bl} composing the stratum. Jackknife estimator of the variance can be defined, then, as follows

$$\hat{V}_{Jack}(\hat{\eta}) = \sum_{h=1}^{H} \frac{n_{Ah}-1}{n_{Ah}} \sum_{i \in s_{Ah}} (\hat{\eta}^A(hi) - \overline{\eta}^{Ah})^2 + \sum_{l=1}^{L} \frac{n_{Bl}-1}{n_{Bl}} \sum_{i \in s_{Bl}} (\hat{\eta}^B(lj) - \overline{\eta}^{Bl})^2, \tag{1.26}$$

where $\hat{\eta}^A(hi)$ is the value taken by $\hat{\eta}$ after dropping unit i of stratum h from sample s_{Ah} and $\overline{\eta}^{Ah}$ is the mean of values $\hat{\eta}^A(hi)$. $\hat{\eta}^B(lj)$ and $\overline{\eta}^{Bl}$ can be defined in a similar way. Again, one can include an approximate finite-population correction in any stratum needing it. 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.

Stratified cluster sampling is a very common design in practice. The jackknife variance estimator when a stratified sample of clusters is selected is now illustrated. Suppose that frame A has H strata and stratum h has N_{Ah} observation units and $\widetilde{N}_{Ah}$ primary sampling units (clusters), of which $\widetilde{n}_{Ah}$ are sampled. Frame B has L strata, and stratum l has N_{Bl} observation units and $\widetilde{N}_{Bh}$ primary sampling units, of which $\widetilde{n}_{Bl}$ are sampled.

To define the jackknife estimator of the variance, let $\widetilde{\eta}^A(hj)$ be the estimator of the same form as $\hat{\eta}$ when the observations of sample primary sampling unit j of stratum h from sample in frame A are omitted. Similarly, $\widetilde{\eta}^B(lk)$ is of the same form as $\hat{\eta}$ when the observations of sample primary sampling unit k of stratum l from sample in frame B are omitted. The jackknife variance estimator is then given by

$$\hat{V}_{Jack}(\hat{\eta}) = \sum_{h=1}^{H} \frac{\widetilde{n}Ah-1}{\widetilde{n}_{Ah}} \sum_{j=1}^{\widetilde{n}_{Ah}} (\widetilde{\eta}^A(hj) - \widetilde{\eta}^{Ah})^2 + \sum_{l=1}^{L} \frac{\widetilde{n}_{Bl}-1}{\widetilde{n}_{Bl}} \sum_{k \in s_{Bl}} (\widetilde{\eta}^B(lk) - \widetilde{\eta}^{Bl})^2, \tag{1.27}$$

where $\widetilde{\eta}^{Ah}$ is the mean of values $\widetilde{\eta}^A(hj)$ and $\widetilde{\eta}^{Bl}$ is the mean of values $\widetilde{\eta}^B(lk)$.

Lohr (2007) proposed two bootstrap variance estimators for dual frame surveys assuming that any generic estimator $\hat{\eta}$ may be expressed through a function, let say h, of the design weights for the two frames. The first variance estimator suggested is called separate bootstrap estimator and it is similar

in form to the jackknife variance estimator (1.25), since the bootstrap is carried out separately in each frame. The separate bootstrap estimator can be defined as

$$\hat{V}_{Boot_S}(\hat{\eta}) = \frac{1}{B_1}\sum_{b=1}^{B_1}(\hat{\eta}^{*A}(b) - \hat{\eta})^2 + \frac{1}{B_2}\sum_{b=1}^{B_2}(\hat{\eta}^{*B}(b) - \hat{\eta})^2, \tag{1.28}$$

where B_1 and $\hat{\eta}^{A*}(b)$ are, respectively, the number of bootstrap iterations in frame A and the bootstrap estimator of η obtained by substituting the original design weights for the bootstrap weights for iteration b only in frame A. B_2 and $\hat{\eta}^{B*}(b)$ are defined similarly for frame B. The number of bootstrap iterations in the frames, B_1 and B_2, may differ and they are determined beforehand by the investigator. In that sense, the bootstrap procedure is more flexible than the jackknife. Another relevant advantage of the bootstrap method compared to the jackknife is than it can be applied to nonsmooth functions (as the median).

To construct the second estimator, denominated combined bootstrap estimator, the resampling is carried out jointly for the whole sample of observations. As a result, the following estimator is formulated:

$$\hat{V}_{Boot_C}(\hat{\eta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\eta}^{*}(b) - \hat{\eta})^2, \tag{1.29}$$

begin, in this case, B the number of resamples drawn and $\hat{\eta}^{*}(b)$ the estimator of η computed by replacing the original design weights for the bootstrap weights for iteration b in both frames.

The three variance estimation procedures (Taylor linearization, jackknife and bootstrap) are easy to apply in practice and they may be extended to the case of three or more frames in a simple way. Nonetheless, bootstrap and, especially, jackknife, require a computation effort that might be heavy even for the nowadays advanced computers.

1.5 Software for estimation in dual frame surveys

Several software packages have been developed to facilitate the analysis of complex survey data and implement some of these estimators as SAS, SPSS, Systat, Stata, SUDAAN or PCCarp. The repository CRAN contains several R packages that include these design-based methods typically used in survey methodology to treat samples selected from one sampling frame (e.g. survey (Lumley, 2014), sampling

(Tillè and Matei, 2012), laeken (Alfons *et al.*, 2014) or TeachingSampling (Gutiérrez Rojas, 2014) among others). Templ (2014) performs a detailed list of packages that includes methods to analyse complex surveys.

However, standard software packages for complex surveys can not be used directly when the sample is obtained from a dual frame survey because the classical design-based estimators are severely biased and there is a underestimation of standard errors. Weighted analyses with standard statistical software, with certain modified weights, can yield correct point estimates of population parameters but still yield incorrect results for estimated standard errors. As exposed in Section 1.3, an important number of authors have developed methods for estimating population means and totals from dual frame surveys but most of these methods require ad-hoc software for their implementation. Unfortunately, there is no software incorporating these estimation procedures for handling dual frame surveys.

1.6 Estimation in three or more frames

Although the majority of the estimators proposed in multiple frames were defined under a dual frame context, for some time now several estimators for the case of three or more sampling frames have been formulated in response to emerging needs in sampling. Indeed, it is clear that the internet has become in a very important data source that offers inexpensive ways to collect information. Couper (2000) analyzes the issues and challenges related with web surveys concluding that this kind of surveys already offer enormous potential for survey researchers which is likely only to improve with time. Within multiple frame context, Lohr (2010) points that web surveys will play a very important role in the future development of multiple frame surveys. So, in the near future it is very likely that dual frame surveys consisting of a cell and a landline frame evolve to multiple frame surveys incorporating a third frame of web users, as represented in Figure 1.6.

As it will be shown, while some of the estimators proposed for a multiple frame setting are the extension of their counterparts in the dual frame context, others have been developed using specific techniques of estimation for three or more sampling frames.

Working in a multiple frame context implies an increase in the complexity of the notation. So, let suppose that $A_1, \ldots, A_q, \ldots, A_Q$ is a collection of $Q \geq 3$ overlapping frames of sizes $N_1, \ldots, N_q, \ldots, N_Q$. As in the dual frame context, all of them can be incomplete but it is assumed that overall they cover the

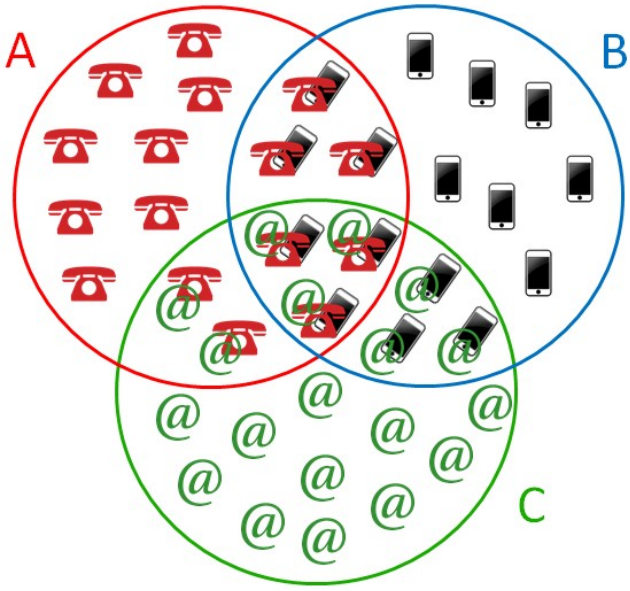


Figure 1.6: A three frame setting composed of a frame of landline users, a frame of mobile phone users and a frame of web users.

entire target population. 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 complementary of a set.

Our aim is, again, to estimate Y, the total of a quantitative variable, which may be expressed as $Y = \sum_{k=1}^{N} y_k$, being y_k the value observed for the $k-th$ individual of the population. The total Y may be rewritten as

$$Y = \sum_{q=1}^{Q} \sum_{k \in U_q} \frac{y_k}{m_k} \tag{1.30}$$

where m_k indicates the number of frames the $k-th$ unit belongs to, i.e. the multiplicity of the $k-th$ unit.

Let s_q be a sample drawn from frame A_q under a particular sampling design d_q, independently for $q = 1, \ldots, Q$, and let π_k^q and π_{kl}^q be the first and second order inclusion probabilities under the sampling design, respectively. Let define $d_k^q = 1/\pi_k^q$ as the sampling weight considered in frame q. Let suppose that n_q is the size of sample s_q and that $s = \cup_q s_q$. When no confusion is possible and for ease of notation, we consider $\pi_k = \pi_k^q$, $\pi_{kl} = \pi_{kl}^q$ and $d_k = d_k^q$ for all sample units k, l such that $k, l \in s_q$.

Lohr (2006) formulated the multiple frame extension of some of the estimators originally proposed

for the dual frame case, as (1.5) or (1.7). As was the case for two frames, the optimal versions of these estimators are asymptotically efficient but they are not internally consistent since they use a different set of weights for each response variable considered. Moreover, they are often unstable in small or moderate samples with more than two frames because the optimal estimated parameters involved in the computation of the estimators are functions of large estimated covariances matrices. They also followed the so called single frame approach used by Kalton and Anderson (1986) to propose a single frame estimator in a multiple frame context. This estimator is in the form:

$$\hat{Y}_{KA} = \sum_{k \in s} y_k d_k^{KA} \tag{1.31}$$

with $d_k^{KA} = \frac{1}{\bar{\pi}_k}$, where $\bar{\pi}_k = \sum_{q' \ni k} \pi_k(q')$.

To compute this estimator it is necessary to know not only the number of frames each unit belongs to but the specific frames the unit is included in. This can be an important drawback specially if misclassification issues are present. The authors also proposed the following pseudo-maximum likelihood estimator for the multiple frame context:

$$\hat{Y}_{PML} = \sum_{k \in s} y_k d_k^{PML}(q) \tag{1.32}$$

where the weights $d_k^{PML}(q)$ can be defined as

$$d_k^{PML}(q) = d_k(q) f(q) \sum_{K: q \in K} \frac{\hat{N}_K \delta_k(K)}{\sum_{j \in K} f(j) \hat{N}_K(j)}$$

with $f(q) = \frac{1}{deff_Y(q)} \frac{n_q}{N_q}$, being $deff_Y(q)$ the design effect for the variable Y in the q-th frame. Values $\hat{N}_K(q)$ can be computed as $\hat{N}_K(q) = \sum_{k \in s_q} d_k(q) \delta_k(K)$, with $\delta_k(K)$ the indicator variable for domain K that takes the value 1 whether $k - th$ individual belongs to domain K and the value 0 otherwise. The estimated domain sizes $\hat{N}_K$ are the solution of a system of non linear equations.

The pseudo maximum likelihood is consistent and usually works well in practical situations but it is complex to compute for a general sampling design, since numerical procedures are required to obtain the values $\hat{N}_K$.

Mecatti (2007) also considered a single frame approach to propose the following estimator

$$\hat{Y}_M = \sum_{k \in s} y_k d_k^M, \tag{1.33}$$

with $d_k^M = \frac{d_k}{m_k}$. The previous estimator, often called single frame multiplicity estimator, only requires the knowledge of the multiplicity of each unit, i.e. the number of frames the unit is included in, no matter which are these frames. This estimator can be adjusted using a raking ratio approach to get a single frame raking ratio multiplicity estimator where a new set of weights, resulting from an iterative procedure, is utilized.

In 2011, Singh and Mecatti proposed a composite multiplicity estimator, which generalizes the single frame multiplicity estimator. This estimator can be written as

$$\hat{Y}_{CM} = \sum_{k \in s} y_k d_k^{CM} \tag{1.34}$$

where

$$d_k^{CM} = \frac{\lambda_k d_k + (1 - \lambda_k) d_k^{KA}}{m_k}$$

with

$$\lambda_k = \frac{\sum_{q' \ni k} (1 - \bar{\pi}_k / \pi_k(q')) \pi_k(q') (1 - \pi_k(q'))}{\sum_{q' \ni k} (1 - \frac{\bar{\pi_k}^2}{\pi_k(q')^2} - \frac{2\bar{\pi_k}}{\pi_k(q')}) \pi_k(q') (1 - \pi_k(q'))}$$

Let suppose now that information about a set of auxiliary variables is available. Let $\check{\mathbf{X}}^q = (X^{q1}, X^{q2}, \ldots, X^{qp_q})'$ be a set of p_q auxiliary variables observed in the $q - th$ frame, so the vector $\mathbf{x}_k^q = (x_k^{q1}, x_k^{q2}, \ldots, x_k^{qp_q})'$ contains the values of the variables $\check{\mathbf{X}}^q$ for the k-th individual of the frame q. Auxiliary variables may differ in each frame, i.e. $\check{\mathbf{X}}^q \neq \check{\mathbf{X}}^r, q, r = 1, \ldots, Q, q \neq r$. For the sample coming from frame q, the values of the variables $(y_k, \mathbf{x}_k^q)$ are observed.

Rao and Wu (2010) followed a single frame multiplicity based approach to extend their pseudo empirical likelihood estimator for the mean of a variable to the multiple frame setting. This estimator can be computed as

$$\hat{\bar{Y}}_{PEL} = \sum_{k \in s} y_k p_k(q) \tag{1.35}$$

with $p_k(q)$ maximizing the likelihood function

$$l_{PEL}(\mathbf{p}_1, \ldots, \mathbf{p}_Q) = \frac{\sum_{q=1}^{Q} n_q}{\sum_{k \in s} d_k^M} \sum_{k \in s} d_k^M log(p_k(q))$$

subject to

$$\sum_{k \in s} p_k(q) = 1$$

$$\sum_{k \in s} p_k(q) \mathbf{x}_k = \bar{\mathbf{X}}$$

being $\bar{\mathbf{X}} = (\bar{X}^1, \bar{X}^2, \ldots, \bar{X}^p)'$ the vector of the population means of variables $\check{\mathbf{X}}^q$, which are supposed to be the same in all frames.

The calibration techniques proposed by Ranalli *et al.* (2015) for the dual frame case, may be easily extended to the multiple frame context. A model calibrated estimator for the case of more than two sampling frames can be defined as

$$\hat{Y}_{CAL} = \sum_{k \in s} y_k d_k^{CAL} \tag{1.36}$$

where d_k^{CAL} are such that minimize

$$\sum_{k \in s} G\left(d_k^{CAL}, d_k^M\right)$$

subject to

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

$$\sum_{k \in s} d_k^{CAL} \mathbf{x}_k^q \delta_k(A_q) = \mathbf{X}^q, \quad q = 1, ..., Q$$

where $\mathbf{X}^q = (X^{q1}, X^{q2}, \ldots, X^{qp_q})'$ is the vector of population totals for the variables $\check{\mathbf{X}}^q$.

The calibration estimator proposed by Elkasabi *et al.* (2015) may be also extended to a multiple frame setting in an easy way. The multiple frame version of the joint calibration estimator has the form

$$\hat{Y}_{JCE} = \sum_{k \in s} y_k d_k^{JCE} \tag{1.37}$$

with $d_k^{JCE} = d_k(1 + \lambda' \mathbf{x}_k)$ and

$$\lambda^{'} = \left(\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} d_k \mathbf{x}_k \right)^{'} \left(\sum_{k \in s} d_k \mathbf{x}_k \mathbf{x}_k^{'} \right)^{-1}$$

To compute this estimator, the same set of auxiliary variables $\grave{\mathbf{X}} = (X_1, X_2, ..., X_p)$ is assumed to be known in all frames.

Chapter 2

Objectives

As evidenced in the previous chapter, the multiple frame sampling theory has experienced a substantial development since its inception in the second half of the twentieth century. Nevertheless, there are still some aspects that require additional attention which motivated the realization of this thesis. The general purpose of the thesis is to further investigate some topics related with multiple frame surveys that have been sparsely addressed so far. This global purpose will be concretized along this chapter through the definition of specific objectives.

In the literature of multiple frames it is possible to find several simulation studies to compare the different estimation methods exposed in Section 1.3 in terms of bias and variance (see e.g. Lohr and Rao (2006), Rao and Wu (2010), Ranalli *et al.* (2015)). However, in very few cases the estimators were applied simultaneously to data coming from a real survey. On the other hand, in surveys it is frequent to find questions where respondents must select one in a series of options, specially in the cases where the survey focuses in public opinion, health or marketing topics. In these situations, the interest lies in estimating the proportions of respondents selecting each possible option. The first aim of this thesis is to apply the estimation methods described so far to the estimation of proportions from data coming from a real dual frame survey, highlighting the issues that can arise and presenting a way to deal with them.

Dual frame surveys are widely used both by statistical agencies and private companies due to their amply proven benefits. One of the main reasons of the recent rise of the popularity of dual frame surveys is the steady increase in the use of telephone surveys, which have replaced all other data collection

methods (the majority of which were face-to-face interviews). In some subject areas (e.g., electoral), face-to-face surveys have been completely ousted by telephone interviewing. Telephone surveys present some drawbacks with regard to coverage, due to the absence of a telephone in some households and the generalized use of mobile phones, which are sometimes replacing fixed (land) lines entirely. Dual frame telephone surveys that combine Random-Digit-Dialing (RDD) landline telephone samples and cell phone samples are a good solution to that issue since they reduce the noncoverage due to cell-only households in RDD landline telephone surveys. Nevertheless, and as noted in Section 1.5, there is no specialized software for analysing data coming from a dual frame survey. This thesis aims the creation of some easy-to-use software to handle dual frame data.

All the estimators described in Section 1.3 were originally proposed to estimate the total or the mean of a continuous variable. Although they may also be used to estimate proportions when the main variable has discrete outcomes, they may provide inconsistent estimates, since estimations over all categories may do not add up to 1, which is desirable in that situations. Therefore, more adequate approaches are required to provide appropriate results. The third objective of this thesis is to propose new estimation techniques to estimate proportions for qualitative response variables.

Chapter 3

Methodology

The breadth of the objectives this thesis pursues makes the use of a variety of techniques to fulfill them.

To reach the first of the objectives, a real phone dual frame survey (considering a frame of landline users and another frame of cell phone users) focused on the opinions of the population regarding immigration in the region of Andalusia (in Spain) has been analyzed. At this point, a first issue related with the sample size allocation arose. Traditionally, in one frame surveys where the sampling frame is composed of landline users, a list including all the individuals of the population is available and, therefore, classical sampling designs as simple random sampling or stratified sampling can be used to select samples. Conversely, when conducting cell phone surveys one does not have a list of the individuals composing the population so alternative methods should be used to select the samples. Among these methods, the random digital dialing (RDD) is one of the most used ones. The issue comes when both frames are sampled simultaneously in a dual frame survey and it is needed to determine the method to draw the samples. Fortunately, dual samples surveys are quite flexible in this aspect since they allow a different data collection procedure in each frame. The key point, then, is to determine the optimal (in some sense) number of individuals from each frame who should be interviewed. In the specific case of this survey, the issue was solved by allocating the predefined global sample size by frames considering a minimum variance criterion taking into account the costs (Pasadas and Trujillo, 2013) and the percentage of possession of each type of device (following Hartley, 1962).

The sampled individuals answered to a selection of questions with discrete outcomes related to im-

migrants and immigration policies. The responses were analyzed and point estimates for proportions using most of the estimators described in Section 1.3 were provided. Since these estimations methods were originally conceived to estimate parameters of continuous variables, estimations of proportions were carried out from the values of a dichotomous variable that was created for each category of each response variable. Therefore, for a category of a given response variable, let consider the dichotomous variable Z_i, so that z_{ki} is the value of Z_i for the $k-th$ individual of the sample. z_{ki} takes the value 1 whether the $k-th$ individual has selected the $i-th$ outcome of the variable and 0 otherwise. Estimated proportion for the $i-th$ category was computed, then, as $\hat{P}_i = \hat{Z}_i/\hat{N} = \sum_{k\in s} z_{ki}/\hat{N}$, being $\hat{N}$ an estimation of the population size.

From the formula used to estimate the proportions it follows that the estimation of the population size, $\hat{N}$, has an important impact on the estimates of the proportions. Thus, it is important to have an accurate estimation of N to achieve good estimations for the proportions. Henceforth, a comprehensive study on the effect on estimation of using different values for the population size extracted from different sizes is carried out.

On the other hand, some of the estimation methods considered involves the estimation of variances and covariances which require second order inclusion probabilities, which were not available in the survey. To overcome this concern, the approximation proposed by Deville (1992) to estimate variances from first order inclusion probabilities is used where needed. According to this approximation, the variance of the estimator of the total of a continuous variable Y may be estimated as

$$\hat{V}(\hat{Y}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s} (1-\pi_k) \left(\frac{y_k}{\pi_k} - \sum_{l\in s} a_l \frac{y_l}{\pi_l} \right)^2 \tag{3.1}$$

where $a_k = (1-\pi_k)/\sum_{l\in s}(1-\pi_l)$.

Results also include interval estimation using the method for variance estimation proposed for each author and jackknife variance estimation, which allow the comparison between estimates.

The objective related to the creation of a software for the analysis of data coming from a dual frame survey has been achieved by the implementation of *Frames2*, a new *R* package for point and interval estimation from dual frame sample data. The development of the package has been carried out taking into account statistical and computational criteria to obtain a comprehensive and efficient software.

Therefore, the functions composing the package have been implemented such that they carry out a

strong argument check to guarantee validity of the arguments and so, to prevent errors when making subsequent computations. Aspects as the presence of missing values in the arguments, the number of main variables observed in the samples (that should match), the length of the arguments in each sample (that should also match) or the values for arguments indicating the domain each unit belongs to (which only can be "a" or "ab" for frame A or "b" or "ba" for frame B) are checked. If any issue is encountered, the function displays an error message indicating the problem and the argument causing it, so that the user can manage errors easily. Furthermore, each function has additional checks depending on its specific characteristics or arguments.

Much attention has also been devoted to computational efficiency. Frequently, populations in a survey are extremely large or it is needed to keep sampling error below a certain value. As a consequence, one needs to consider large sample sizes, often in the order of tens of thousands sampling units. In these situations, computational efficiency of functions is essential, particularly when several variables are considered. Otherwise, user can face high runtimes and heavy computational loads. In this sense, functions of *Frames2* are developed according to strict efficiency measures, using the power of R to the matrix calculation to avoid loops and increase the computational efficiency.

Functions of *Frames2* have been implemented from an user-oriented perspective to increase usability. In this sense, most input parameters (which are the communication channel between the user and the function) are divided into two groups, depending on the frame they come from. This is to adapt functions as much as possible to the usual estimation procedure, in which the first step is to draw two independent samples, one from each frame. On the other hand, estimation details are managed internally by functions so that they are not visible for the user, who does not need to manage them.

Construction of functions has been carried out so that they perform properly in as many situations as possible. As noted in introductory section, one can face several situations when using two sampling frames depending on their relative positions. Although the most common situation is the one depicted in Figure 1.1, cases shown in Figures 1.2 and 1.3 may arise as well. All estimators described but PEL can be modified to cover these three situations, so corresponding functions of *Frames2* include necessary changes to produce estimates irrespective of the situation.

On the other hand, it is usual, when conducting a survey, to collect information on many variables of interest. To adapt to such situations, all functions are programmed to produce estimates when there are more than one variable of interest with only one call. To this end, parameters containing information

about main variables observed in each frame can be either vectors, when only one variable is considered or matrices or data frames, when there are several variables under study. Cases in which the main aim of the survey is the estimation of population means or proportions are also very frequent. Hence, from the estimation of the population total for a variable, functions compute estimation of the mean as $\hat{\bar{Y}} = \hat{Y}/\hat{N}$. To obtain the estimation of the population size, functions internally apply the estimation procedure at issue to indicator vectors $\mathbf{1}_A$ and $\mathbf{1}_B$ of sizes n_A and n_B, respectively.

To get maximum flexibility, functions have been programmed to calculate estimates in cases in which user disposes of first and second order inclusion probabilities and in those other situations in which only first order ones are available, indistinctly. Variance estimations from only first order inclusion probabilities are obtained by applying Deville's method (3.1), when needed.

Finally, to reach the third objective, appropriate models to deal with discrete response variable are considered. Firstly, let assume that data from respondents who provide a single choice from a list of non ordered alternatives, coded as $1, 2, \ldots, m$ are collected. 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, let consider the class of indicators Z_i, $i = 1, \ldots, m$, defined previously. These indicators are such that, for each unit $k \in U$, $z_{ki} = 1$ if $y_k = i$ and $z_{ki} = 0$ otherwise. The problem thus, is to estimate the proportions

$$P(Y = i) = P_i = \frac{1}{N} \sum_{k \in U} z_{ki}, \quad i = 1, 2, \ldots, m \tag{3.2}$$

Such proportions are such that

$$P_i = \frac{1}{N}(Z_{ai} + \theta Z_{abi} + (1 - \theta) Z_{abi} + Z_{bi}), \tag{3.3}$$

where $\theta \in (0, 1)$ and $Z_{ai} = \sum_{k \in a} z_{ki}$, $Z_{abi} = \sum_{k \in ab} z_{ki}$ and $Z_{bi} = \sum_{k \in b} z_{ki}$.

As noted before, auxiliary information is often available in survey sampling and may be used to obtain more accurate estimators. Then, suppose that values $\mathbf{x}_k$ of auxiliary variables $\check{X}$ are known for each $k \in U$. Moreover, the distribution, or at least some summary statistics, of these auxiliary variables in the population are supposed to be known. Let assume that the population under study $\mathbf{y} = (y_1, ..., y_N)^T$

is the determination of a set of super-population random variables $\mathbf{Y} = (Y_1, ..., Y_N)^T$ s.t.

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

that is, a multinomial logistic model is used to relate the main response and the auxiliary variables. Let $\boldsymbol{\beta}$ be the parameter vector $(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_m^T)^T$. As a first step, estimation of the superpopulation parameter $\boldsymbol{\beta}$ using the sample data should be considered. This estimation will be carried out in a different way depending on the auxiliary information available, resulting in two groups of estimators: a first group, composed of 4 estimators, where the same set of auxiliary variables for all population units is assumed and a second one, which includes 2 estimators, where the auxiliary variables differ by frame.

To create the first group of estimators, let assume that, for each unit in the population, information about one vector of auxiliary variables $\check{\mathbf{X}}$ is known. In this case, for each unit $k \in U$ the value of $\mathbf{x}_k$ is available and, for each unit $k \in s$, the value of the main variable y_k is also observed. Parameter $\boldsymbol{\beta}$ may be estimated by maximizing the π-weighted log-likelihood function given by

$$\ell_{d^\circ}(\boldsymbol{\beta}) = \sum_{i=1,...,m} \left(\sum_{k \in s_A} d_k^\circ z_{ki} \ln \mu_{ki} + \sum_{k \in s_B} d_k^\circ z_{ki} \ln \mu_{ki} \right), \tag{3.4}$$

where the weights d° are the ones defined in (1.2). Given the estimate $\widehat{\boldsymbol{\beta}^\circ}$ of $\boldsymbol{\beta}$, the following estimates for μ_{ki} may be defined:

$$p_{ki}^o = \widehat{\mu}_{ki} = \frac{\exp(\mathbf{x}_k^T \widehat{\boldsymbol{\beta}}_i^o)}{\sum_{r=1,...,m} \exp(\mathbf{x}_k^T \widehat{\boldsymbol{\beta}}_r^o)}. \tag{3.5}$$

Since the vector $\mathbf{x}_k$ is known for all units of the population U, the values p_{ki}^o are available for all $k \in U$.

An alternative way of estimating $\boldsymbol{\beta}$ is maximizing the π-weighted log-likelihood

$$\ell_{\tilde{d}}(\boldsymbol{\beta}) = \sum_{i=1,...,m} \sum_{k \in s} \tilde{d}_k z_{ki} \ln \mu_{ki}, \tag{3.6}$$

which is similar to (3.4) but using weights $\tilde{d}$ (defined in (1.3)) instead of d°. In that case, the resulting

estimate, $\widehat{\tilde{\boldsymbol{\beta}}}$, may be used to compute the following estimations for the individual probabilities

$$\tilde{p}_{ki} = \widehat{\mu}_{ki} = \frac{\exp(\mathbf{x}_k^T \widehat{\tilde{\boldsymbol{\beta}}}_i)}{\sum_{r=1,\ldots,m} \exp(\mathbf{x}_k^T \widehat{\tilde{\boldsymbol{\beta}}}_r)}. \tag{3.7}$$

On the other hand, to formulate the second group of estimators let suppose that a different set of variables is known in each frame, that is, values $\mathbf{x}_k^A$ of the vector of auxiliary variables $\check{\mathbf{X}}^A$ are known for all the units composing frame A and values $\mathbf{x}_k^B$ of another vector $\check{\mathbf{X}}^B$ are known for the units included in frame B. In that case, a different model should be considered in each frame to properly represent the relationship between the auxiliary variables and the main response. Then, For each $k \in A$, values of the auxiliary vector $\mathbf{x}_k^A$ are known and, thus, we may compute the probabilities

$$p_{ki}^A = \frac{\exp(\mathbf{x}_k^{AT} \widehat{\boldsymbol{\beta}}_i^A)}{\sum_{r=1,\ldots,m} \exp(\mathbf{x}_k^{AT} \widehat{\boldsymbol{\beta}}_r^A)} \tag{3.8}$$

where we estimate $\boldsymbol{\beta}^A$ by maximizing $\ell_{d^A}(\boldsymbol{\beta}^A) = \sum_{i=1,\ldots,m} \sum_{k \in s_A} d_k^A z_{ki} \ln \mu_{ki}$ using the sample data from s_A. Sample s_A includes, together with values of the auxiliary variables, the values of the main response y_k (and, therefore of z_{ki}). Similarly we obtain p_{ki}^B for $k \in B$.

Estimates p_{ki}^o, $\tilde{p}_{ki}$ and p_{ki}^A and p_{ki}^B may be used as auxiliary information to define estimators.

As may be noted, this estimation approach has been exposed in a general context which is barely affected by the number of sampling frames involved. This indicates that, although the dual frame case is usually the starting point for the estimation in a multiple frame setup, this methodology can be easily extended to the case where three or more frames are available for sampling. Good evidence of this can be found in the fact that a general multiple frame context, with $Q \geq 3$ frames, has been considered when studying response variables whose categories may be somehow ordered in Appendix 4.

Analysis of responses variables with ordered outcomes is carried out following a similar approach than the one used for variables with non ordered categories.

Considering the same multiple frame setup exposed in Section 1.6, let consider the discrete survey variable y to represent the choice of the respondents from a list of ordered alternatives. We code these alternatives as $1, 2, \ldots, m$, with $1 < 2 < \cdots < m$. Therefore, y is an m-valued survey variable with y_k the value observed for the $k - th$ individual of the population. The objective is to estimate proportions

(3.2), which can be rewritten as follows

$$P(Y=i)=P_i=\frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in U_q}\frac{z_{ki}}{m_k},\quad i=1,2,\ldots,m \tag{3.9}$$

where m_k indicates the number of frames the $k-th$ unit belongs to, i.e. the multiplicity of the $k-th$ unit and z_{ki} are, again, the values of the indicator variables $Z_i, i=1,...,m$, such that for each unit $k\in U$ $z_{ki}=1$ if $y_k=i$ and $z_{ki}=0$ otherwise.

Let suppose that information about auxiliary variables is available. Let $\check{X}_q=(\mathbf{x}_{q1},\mathbf{x}_{q2},\ldots,\mathbf{x}_{qp_q})^T$ be a set of p_q auxiliary variables observed in the q-th frame, so the vector $\mathbf{x}_{qk}=(x_{q1k},x_{q2k},\ldots,x_{qp_qk})^T$ contains the values of the variables $\boldsymbol{x}_q$ for the k-th individual of the frame q. Auxiliary variables may differ in each frame, i.e. $\mathbf{X}_q\neq\mathbf{X}_r, q,r=1,\ldots,Q, q\neq r$, so the most general and realistic situation is considered. For the sample coming from frame q, the values of the variables $(y_k,\mathbf{x}_{qk})$ are observed.

Taking into account the ordinal nature of the response variable, an ordinal model should be considered instead of a multinomial one to properly relate the main and the auxiliary variables. Therefore, in frame q, the finite population under study $\mathbf{y}=(y_1,...,y_N)^T$ is the determination of the superpopulation random variable vector $\mathbf{Y}=(Y_1,...,Y_N)^T$ s.t.

$$\mu_i^q(\mathbf{x}_{qk})=P(Y_k=i|\mathbf{x}_{qk})=E(Z_{ki}|\mathbf{x}_{qk})=\begin{cases}\frac{\exp(\alpha_i^q+\boldsymbol{\beta}_i^q\mathbf{x}_{qk})}{1+\exp(\alpha_i^q+\boldsymbol{\beta}_i^q\mathbf{x}_{qk})} & ,i=1\\ \frac{\exp(\alpha_i^q+\boldsymbol{\beta}_i^q\mathbf{x}_{qk})}{1+\exp(\alpha_i^q+\boldsymbol{\beta}_i^q\mathbf{x}_{qk})}-\frac{\exp(\alpha_{i-1}^q+\boldsymbol{\beta}_{i-1}^q\mathbf{x}_{qk})}{1+\exp(\alpha_{i-1}^q+\boldsymbol{\beta}_{i-1}^q\mathbf{x}_{qk})} & ,i=2,...,m\end{cases}. \tag{3.10}$$

assuming that Y_k are conditionally independent given $\mathbf{x}_{qk}$. An important property that is usually supposed to be accomplished when working with ordinal models is the proportional odds property. According to this property, effects of the predictors are the same across all the categories. This implies that $\boldsymbol{\beta}$ parameters associated to independent variables are fixed and independent of the category considered, so constraints of the superpopulation model can be rewritten as

$$\mu_i^q(\mathbf{x}_{qk})=P(Y_k=i|\mathbf{x}_{qk})=E(Z_{ki}|\mathbf{x}_{qk})=\begin{cases}\frac{\exp(\alpha_i^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})}{1+\exp(\alpha_i^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})} & ,i=1\\ \frac{\exp(\alpha_i^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})}{1+\exp(\alpha_i^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})}-\frac{\exp(\alpha_{i-1}^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})}{1+\exp(\alpha_{i-1}^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})} & ,i=2,...,m\end{cases} \tag{3.11}$$

The proportional odds property provides more parsimonious models which are, therefore, easier to implement and interpret. As with the multinomial model, population parameters $\alpha_i, i=1,...m$ and $\boldsymbol{\beta}$ are

generally unknown and must be estimated from the sample data. Considering, again, a maximum likelihood approach we can obtain the estimates for the $\boldsymbol{\theta}^q$-parameter $\boldsymbol{\theta}^q = (\alpha_1^q, \ldots, \alpha_m^q, \boldsymbol{\beta}^q)$ by maximizing the following loglikelihood function

$$\ell(\boldsymbol{\theta}^q) = \sum_{i=1,\ldots,m} \sum_{k \in s_q} d_k^q z_{ki} \ln \mu_i^q(\mathbf{x}_{qk}), \tag{3.12}$$

and we denote it by $\hat{\boldsymbol{\theta}}^q = (\hat{\alpha}_1^q, \ldots, \hat{\alpha}_m^q, \hat{\beta}^q)$. Using these maximum likelihood estimates, we can define an estimator for probabilities for each category as follows:

$$p_{ki}^q = \hat{\mu}_i^q(\mathbf{x}_{qk}) = \begin{cases} \frac{exp(\hat{\alpha}_i^q + \hat{\beta}^q \mathbf{x}_{qk})}{1+exp(\hat{\alpha}_i^q + \hat{\beta}^q \mathbf{x}_{qk})} & , i = 1 \\ \frac{exp(\hat{\alpha}_i^q + \hat{\beta}^q \mathbf{x}_{qk})}{1+exp(\hat{\alpha}_i^q + \hat{\beta}^q \mathbf{x}_{qk})} - \frac{exp(\hat{\alpha}_{i-1}^q + \hat{\beta}^q \mathbf{x}_{qk})}{1+exp(\hat{\alpha}_{i-1}^q + \hat{\beta}^q \mathbf{x}_{qk})} & , i = 2, ..., m \end{cases} \tag{3.13}$$

Alternatively to (3.12), model parameters for the q-th frame can be estimated maximizing the following loglikelihood function

$$\ell(\boldsymbol{\theta}^q) = \sum_{i=1,\ldots,m} \sum_{k \in s_q} d_k^M z_{ki} \log \mu_i^q(\mathbf{x}_{qk}), \tag{3.14}$$

yielding to the probability estimates

$$p_{ki}^{\star q} = \hat{\mu}_i^q(\mathbf{x}_{qk}) = \begin{cases} \frac{exp(\hat{\alpha}_i^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})}{1+exp(\hat{\alpha}_i^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})} & , i = 1 \\ \frac{exp(\hat{\alpha}_i^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})}{1+exp(\hat{\alpha}_i^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})} - \frac{exp(\hat{\alpha}_{i-1}^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})}{1+exp(\hat{\alpha}_{i-1}^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})} & , i = 2, ..., m \end{cases} . \tag{3.15}$$

As in the multinomial case, both sets of estimates (3.13) and (3.15) may be used in the definition of estimators for proportions of ordinal responses variables.

Chapter 4

Results

Some important results have been derived from the research carried out in this thesis. The most noticeable ones are summarized below.

From the analysis of the opinion survey about immigrants and immigration policies performed in Appendix 1 some aspects may be highlighted:

- There are no important differences between the estimates produced with the single frame or dual frame approach.
- Among all the estimation strategies, the calibration method performs best and produces the smallest confidence interval.
- The jackknife method often produces better intervals than methods based on the estimated variance given by the authors (except for the pseudo empirical likelihood intervals).

Results obtained show a negative view towards immigration that continues to spread. In the moment of the data collection, 59-61% of the individuals surveyed in Andalusia stated that immigration is bad or very bad for the region (in the previous edition of the study, in 2011, the corresponding figure was 58 %, and in the first such survey, in 2005, it was only 51%). Perceptions regarding the number of immigrants, however, have changed in the opposite direction: there is now a lower percentage of people who say there are too many immigrants (from 51 % in 2011 to current levels of 40-42 %), while the other scores have risen slightly.

Another important result derived from the thesis is the software *Frames2*. *Frames2* is a new R package for point and interval estimation from dual frame sampling. The initial version consisted of eight main functions (`Hartley`, `FB`, `BKA`, `SFRR`, `PML`, `PEL`, `CalSF` and `CalDF`), implementing most of the estimators described in Section 1.3. The package also includes an additional function called `Compare` which provides a summary with all possible estimators that can be computed from the information provided as input. Moreover, six extra functions implementing auxiliary operations, like computation of Horvitz-Thompson estimators or of the covariance between two Horvitz-Thompson estimators, have also been included in the package to achieve a more understandable code. Finally, the package includes eight more functions, one for each estimator, for the calculation of confidence intervals based on the jackknife variance estimator.

The package is freely available at the CRAN repository following the URL `https://cran.r-project.org/web/packages/Frames2/index.html`. In that web site one may also find a reference manual including information about all the functions composing the package and some vignettes illustrating how to use it in different contexts.

On the other hand, a number of estimators for dealing with multinomial response variables in dual frame surveys has been proposed. As specified in the previous section, different set of estimated probabilities (p_{ki}^o, $\tilde{p}_{ki}$ or p_{ki}^A and p_{ki}^B) have been defined, depending on the available auxiliary information. Whatever the case, these probabilities represent the true relationship between the auxiliary variables and the main response.

From probabilities p_{ki}^o defined in (3.5) two estimators are formulated to estimate proportions defined in (3.3) considering dual frame and single frame approaches. The first one is expressed as

$$\widehat{P}_{MLi}^{DW} = N^{-1}\left(\sum_{k\in U} p_{ki}^o + \sum_{k\in s_A} d_k^\circ(z_{ki} - p_{ki}^o) + \sum_{k\in s_B} d_k^\circ(z_{ki} - p_{ki}^o)\right) \tag{4.1}$$

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

Following a model calibrated approach (here, subscript MLC refers to Multinomial-Logistic and

Calibration), a second estimator using probabilities p^o_{ki} is defined. It has the form

$$\widehat{P}^{DW}_{MLCi} = N^{-1}(\sum_{k\in s_A} w^\circ_k z_{ki} + \sum_{k\in s_B} w^\circ_k z_{ki}) \tag{4.2}$$

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:

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

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

Following a similar approach, but considering estimates $\tilde{p}_{ki}$ defined in (3.7) as auxiliary variables, equivalent estimators to (4.1) and (4.2) are defined as

$$\widehat{P}^{SW}_{MLi} = N^{-1}\left(\sum_{k\in U} \tilde{p}_{ki} + \sum_{k\in s_A} \tilde{d}_k(z_{ki} - \tilde{p}_{ki}) + \sum_{k\in s_B} \tilde{d}_k(z_{ki} - \tilde{p}_{ki})\right) \tag{4.3}$$

and

$$\widehat{P}^{SW}_{MLCi} = N^{-1}(\sum_{k\in s_A} \tilde{w}_k z_{ki} + \sum_{k\in s_B} \tilde{w}_k z_{ki}), \tag{4.4}$$

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:

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

Here, the superscript SW stands Single frame setting and auxiliary information available from the Whole population. Again, subscript ML stands Multinomial-Logistic while MLC stands for Multinomial-Logistic and Calibration.

The four estimators (4.1), (4.2), (4.3) and (4.4) have the common characteristic of being defined from a common set of auxiliary variables whose values are available for the whole population. Nevertheless,

different sets of auxiliary variables in each frame may be considered, as noted in previous chapter. In that situations, estimated probabilities p_{ki}^A defined in (3.8) and their counterparts p_{ki}^B may be used to define the following estimators:

$$\widehat{P}_{MLi}^{DF} = N^{-1}\left(\sum_a p_{ki}^A + \eta\sum_{ab} p_{ki}^A + (1-\eta)\sum_{ba} p_{ki}^B + \sum_b p_{ki}^B + \right.$$

$$+\sum_{s_a}(z_{ki}-p_{ki}^A)d_{Ak} + \eta\sum_{s_{ab}}(z_{ki}-p_{ki}^A)d_{Ak} +$$

$$\left.+(1-\eta)\sum_{s_{ba}}(z_{ki}-p_{ki}^B)d_{Bk} + \sum_{s_b}(z_{ki}-p_{ki}^B)d_{Bk}\right).$$

and

$$\widehat{P}_{MLCi}^{DF} = N^{-1}(\sum_{k\in s_A} w_k^\star z_{ki} + \sum_{k\in s_B} w_k^\star z_{ki}) = N^{-1}(\sum_{k\in s} w_k^\star z_{ki}), \tag{4.5}$$

where weights $w_k^\star$ are such that

$$\min \sum_{k\in s_A} G(w_k^\star, d_{Ak}) + \sum_{k\in s_B} G(w_k^\star, d_{Bk}) \qquad \text{s.t.}$$

$$\sum_{k\in s_A} w_k^\star p_{ki}^A = \sum_{k\in a} p_{ki}^A + \eta\sum_{k\in ab} p_{ki}^A,$$

$$\sum_{k\in s_B} w_k^\star p_{ki}^B = (1-\eta)\sum_{k\in ba} p_{ki}^B + \sum_{k\in b} p_{ki}^B,$$

$$\sum_{k\in s_a} w_k^\star = N_a, \quad \sum_{k\in s_b} w_k^\star = N_b,$$

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

Performance of the 6 estimators has been check through different simulations studies resulting in negligible biases and important efficiency gains with respect to customary estimators not using auxiliary information (as (1.8)) and estimators using the auxiliary information through linear models (as (1.15) or (1.18)). Moreover, important length reductions in jackknife confidence intervals respect to estimator (1.8) are obtained when applying the proposed estimators to data coming from a real survey.

Finally, estimated probabilities defined in (3.15) have been used to formulate some estimators for the

proportions of an ordinal response variable in a multiple frame context. First of all, the following two model assisted estimators have been proposed:

$$\hat{P}_{MA1i} = \frac{1}{N}\left(\sum_{q=1}^{Q}\sum_{k\in U}\frac{p_{ki}^{q}}{m_k} - \sum_{k\in s} p_{ki}^{q} d_k^M + \sum_{k\in s} z_{ki} d_k^M\right), \quad i = 1, ..., m \tag{4.6}$$

$$\hat{P}_{MA2i} = \frac{1}{N}\left(\sum_{q=1}^{Q}\sum_{k\in U}\frac{p_{ki}^{q}}{m_k} - \frac{N}{\hat{M}}\sum_{k\in s} p_{ki}^{q} d_k + \sum_{k\in s} z_{ki} d_k^M\right), \quad i = 1, ..., m \tag{4.7}$$

with $\hat{M} = \sum_{k\in s} d_k$. To formulate both estimators we have adapted the approach used by Lehtonen and Veijanen (1998a) 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 in estimator $\hat{P}_{MA1i}$ are weighted by multiplicities m_k to avoid overestimation issues. For this same reason, weights d_k^M are used in the sample sums. Such weighing is intended to make the estimator consistent in the sense that its categories add up to 1. Estimator $\hat{P}_{MA2i}$ is very similar to $\hat{P}_{MA1i}$, with the only difference of using original design weights d_k in one of the sample sums. Due to this, and to ensure the consistency of the estimator, adjustment factor $N/\hat{M}$ is used.

Using probabilities p_{ki}^q as auxiliary variables and considering a model calibration approach, the following estimator may be formulated:

$$\hat{P}_{MC1i} = \frac{1}{N}\sum_{k\in s}\frac{w_k^{\circ}}{m_k} z_{ki}, \quad i = 1, ..., m, \tag{4.8}$$

where weights w_k° are chosen so that they minimize $\sum_{k\in s} G\left(w_k^{\circ}, d_k\right)$, subject to

$$\sum_{k\in s}\frac{w_k^{\circ}}{m_k}\delta_k(A_q) = N_q, \quad q = 1, ..., Q$$

$$\sum_{k\in s}\frac{w_k^{\circ}}{m_k} p_{ki}^{q}\delta_k(A_q) = \sum_{k\in U} p_{ki}^{q}\delta_k(A_q), \quad q = 1, ..., Q, \quad i = 1, ..., m.$$

In the first group of Q calibration constraints, regarding frame sizes, multiplicities m_k are used to properly weight indicator variables $\delta_k(A_q)$ and so, to cancel any overestimation problem. The same reasoning may be applied to the second group of constraints, where the auxiliary variables are also weighted by m_k.

A calibration approach may be considered also when estimates (3.15) are used as auxiliary information.

Similarly to (**??**), another model calibrated estimator may be defined as

$$\hat{P}_{MC2i} = \frac{1}{N} \sum_{k \in s} w_k^\star z_{ki}, \quad i = 1, ..., m \tag{4.9}$$

where, in this case, the weights $w_k^\star$ are such that they minimize $\sum_{k \in s} G\left(w_k^\star, d_k^M\right)$ subject to

$$\sum_{k \in s} w_k^\star \delta_k(A_q) = N_q, \quad q = 1, \ldots, Q$$

$$\sum_{k \in s} w_k^\star p_{ki}^{\star q} \delta_k(A_q) = \sum_{k \in U} p_{ki}^{\star q} \delta_k(A_q), \quad q = 1, \ldots, Q, \quad i = 1, ..., m.$$

Unlike those in $\hat{P}_{MC1i}$, constraints for this calibration estimator do not involve multiplicities. Over-estimation issues are eliminated, then, by considering d_k^M (which are already weighted by m_k) as the starting weights for the calibration. Therefore, resulting weights $w_k^\star$ should be near to those starting weights so they already take into account the multiplicity while still fulfilling the calibration constraints.

The proposed estimators have shown a good behaviour in terms in bias (which may be considered as negligible) and in terms of efficiency gain with respect to customary multiple frame estimators (as (1.33) and (1.36)) in the comprehensive simulation studies carried out. Moreover, proposed estimators work well when applied to real data coming from a dual frame survey.

Chapter 5

Conclusions

Let us remember that the main objective of this thesis is to further study some aspects of the multiple frame methodology that had not been addressed so far. As a result, a number of estimators for proportions of discrete response variables have been proposed. Furthermore, software for the analysis of data coming from dual frame surveys has been released. The main findings derived of the analysis of the results obtained are detailed below.

Implementation of multiple frame surveys may be challenging in some cases due to the increase of the complexity with respect to surveys considering only one frame. Focusing on dual frame surveys, different approaches are available for the analysis of data coming from that kind of surveys. The screening approach is quite interesting since it allows the use of the well known techniques for stratified samples. However screening is barely applied due to, in most situations, duplicated units of the overlap domain can not be identified, which is fundamental in this technique. A dual frame or a single frame methodology should be, then, considered.

The application of the customary estimators to the data coming from a real dual frame survey focused on immigration topics allowed their comparison. Calibration, fixed weight, and pseudo maximum likelihood estimators all give internal consistency (which is a desirable property in an estimator), since the same set of adjusted weights is used for all variables. Moreover, in the application, good results were obtained with these procedures. With repeated surveys, the simplicity and transparency of a fixed-weight estimator may be preferred. Fixed-weight adjustments may make year-to year comparisons easier in an

annual survey, where the domain proportions are relatively constant over time. Fixed-weight estimators are also more amenable to weight adjustments for non-response and domain misclassification.

Estimators based on the single frame approach are also very appealing. In addition to being internally consistent, it is shown that they generally provide quite good results when applied to practical situations. They are also quite easy to implement. Nevertheless, single frame estimators present the main drawback of needing extra information regarding the inclusion probabilities of the units belonging to the overlap domain which is not always available making the computation of these estimators impossible to carry out.

On the other hand, variance estimation is a tricky issue when dual frame estimators are used. Resampling methods such as jackknife, which is easy to compute an provide accurate estimates, is advisable to estimate variances. Jackknife constitutes an unifying approach that allows the comparison between estimates of the variance of different estimators.

The use of auxiliary information, which is often available in surveys, may become a double-edged sword. While it is true that "good" auxiliary information (in the sense of well related with the main variable) may improve estimates considerably, poor auxiliary variables may lead to incorrect estimates and confidence intervals too wide to the extent that it would be preferable not to use them in the estimation process. Special care should be taken in at this point.

Variables with discrete outcomes, very common in surveys, should be treated in a special way to get appropriate results. It is important to consider appropriate estimation techniques depending on the nature of the variable of study to get the best results possible. As an example, simulation results carried out in a dual frame setup show that, ordinal estimators presented in the Appendix 4 provide much better results for the proportions of an ordinal response variable than the ones we obtain by applying the multinomial estimators proposed in the third Appendix. In turn, in that situation, results of multinomial estimators are better than results of customary dual frame estimators described in Section 1.3. Both groups, multinomial and ordinal, of proposed estimators require the knowledge of the values of auxiliary variables for each individual in the population, which can be quite a restrictive assumption. This assumption may be somehow relaxed 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 this context, it is not necessary to have the values of the auxiliary variables for the complete list of individuals but only the population count in the multi-way

contingency table, information that can be easily found in databases of national statistical organisms.

Finally, the need for a software for analyzing the data from dual frame surveys led to the release of the R package *Frames2*. The package allows a comprehensive analysis of dual frame data through user-friendly functions. These functions have been implemented following strict criteria regarding computational efficiency to provide results quickly minimizing the computational load. Last version of the package, as well as documentation and illustrative examples on it use may be freely accessed through the URL `https://cran.r-project.org/web/packages/Frames2/index.html`.

Chapter 6

Conclusiones

Recordemos que el objetivo principal que se pretende alcanzar con esta tesis es el estudio en profundidad de algunos aspectos de la metodología de encuestas con marcos múltiples que aún no habían sido tratados. Para alcanzar este objetivo, se ha formulado un buen número de estimadores para las proporciones de variables de respuesta discreta. Del mismo modo, se ha creado un programa para el análisis de datos procedentes de encuestas con marcos duales. Las principales conclusiones que se derivan del análisis de los resultados obtenidos se detallan a continuación

La puesta en práctica de encuestas con marcos múltiples puede suponer un reto en algunos casos debido al aumento de su complejidad en comparación con las encuestas de un único marco. Centrándonos en el caso de encuestas con marcos duales, existen diferentes procedimientos para el análisis de datos provenientes de ellas. La metodología "screening" resulta muy interesante ya que permite el uso de las ampliamente conocidas técnicas de muestreo estratificado para el análisis de los datos. Sin embargo, el "screening" raras veces puede aplicarse debido a que requiere la identificación de las unidades duplicadas en el dominio de solapamiento, lo cual es imposible en muchos casos. En su lugar, debe considerarse una metodología "single frame" o "dual frame".

La aplicación de los estimadores tradicionales para encuestas con marcos duales a los datos procedentes de la encuesta de inmigración posibilitó las comparaciones entre ellos. El de calibración, el de pesos fijos o el de pseudo máxima verosimilitud son estimadores consistentes en el sentido de que utilizan el mismo conjunto de pesos ajustados para la estimación de todas las variables (lo cual es una propiedad deseable

en un estimador). Por otra parte, en la aplicación, estos estimadores proporcionaron buenos resultados. Cuando las encuestas se llevan a cabo de forma periódica, la simplicidad y la transparencia del estimador de pesos fijos pueden resultar decisivos a la hora de decantarse por el uso de este estimador. Además, por ejemplo, el uso del estimador de pesos fijos hace que la comparación entre encuestas que se realizan de forma anual resulte muy sencilla, ya que es habitual que estas encuestas presenten un reparto de la población entre los dominios que componen los marcos bastante constante a lo largo del tiempo. Otra ventaja de este estimador es que puede ser fácilmente ajustado para corregir problemas de no repuesta y de clasificación incorrecta de unidades en dominios.

Los estimadores basados en un enfoque "single frame" también son muy atractivos. Además de ser internamente consistentes, se ha demostrado que funcionan bastante bien cuando se aplican en situaciones reales. También presentan la ventaja de ser sencillos de implementar. No obstante, los estimadores basados en esta metodología muestran el inconveniente de necesitar información adicional relativa a las probabilidades de inclusión de las unidades que pertenen al dominio de intersección, la cual no está siempre disponible, haciendo que estos estimadores no puedan ser calculados en algunas situaciones.

La estimación de la varianza es un aspecto complicado para los estimadores en marcos duales. En estos casos, se recomienda estimar las varianzas mediante el uso de algún método de remuestreo como el jackknife, el cual es sencillo de aplicar y proporciona estimaciones bastante precisas. Además, el jackknife constituye un enfoque unificador que permite la comparación entre las estimaciones de la varianza de distintos estimadores.

El uso de información auxiliar, de la cual se dispone habitualmente en las encuestas, puede convertirse en un arma de doble filo. Si bien es cierto que información auxiliar "buena" (en el sentido de que estar altamente relacionada con la variable de interés) puede mejorar considerablemente las estimaciones, una información auxiliar pobre puede desembocar en estimaciones incorrectas y en intervalos de confianza demasiado amplios. Tanto es así que, en ocasiones, puede ser preferible no considerar la información auxiliar en el proceso de estimación. Por lo tanto, debe prestarse una especial atención en la selección de las variables que se utilizan como auxiliares.

Las variables de respuesta discreta, muy frecuentes en las encuestas, deben tratarse de forma especial si se quieren obtener resultados correctos. Es muy importante aplicar las técnicas de estimación adecuadas en función de la naturaleza de la variable respuesta para así obtener los mejores resultados posibles. A modo de ejemplo, un estudio de simulación realizado considerando dos marcos muestrales mostró que,

dada una variable respuesta con categorías ordenadas, las estimaciones para las proporciones de estas categorías que proporcionaron los estimadores ordinales que se describen en el Apéndice 4 fueron mucho mejores que los resultados proporcionados por los estimadores multinomiales del tercer apéndice. A su vez, en este mismo contexto, las estimaciones obtenidas con los estimadores multinomiales se mostraron mucho más precisas que aquellas resultantes de aplicar los estimadores tradicionales para marcos duales que se describieron en la Sección 1.3. Los dos grupos de estimadores propuestos, multinomiales y ordinales, necesitan conocer los valores de las variables auxiliares para todos los individuos de la poblacion, lo cual puede suponer una limitación importante para su uso. Esta hipótesis puede relajarse cuando se utilizan variables categóricas (como el género o el estado profesional del individuo) o variables cuantitativas categorizadas (como la edad del individuo, agrupada en clases) como información auxiliar. En estos casos, no es necesario disponer de los valores de las variables auxiliares para todos los individuos de la población sino únicamente de las frecuencias poblacionales que aparecen en la tabla de contingencia que recoge los cruces entre las categorías de las variables. Esta información puede extrarse fácilmente de las bases de datos que los organismos nacionales de estadística tienen a disposición del público.

Por último, la necesidad de un software para el análisis de datos procedentes de encuestas con marcos duales resultó en la creación del paquete de R *Frames2*. El paquete permite un completo análisis de datos de encuestas con marcos duales a través del uso de funciones muy sencillas de utilizar para el usuario. Estas funciones se han implementando siguiendo criterios muy estrictos en cuanto a la eficiencia computacional para que proporcionen resultados en el menor tiempo posible minimizando también la carga computacional. La última versión del paquete, así como su manual de uso y ejemplos ilustrativos puede descargarse de forma gratuita en la URL `https://cran.r-project.org/web/packages/Frames2/index.html`.

Chapter 7

Current Research Lines

This thesis explores some aspects of the multiple frame approach which required further investigation. Nevertheless, there are still some points that need additional attention. The study of these topics would suppose a natural extension of this thesis. A brief summary of some of the topics that are currently under investigation is presented below.

- As noted in previous chapters, a number of estimators have been proposed so far to estimate parameters of quantitative variables in a multiple frame context following different approaches, as calibration or likelihood. Nevertheless, there are a noticeable number of estimation techniques which has been applied to single frame surveys but whose performance has not been evaluated in a multiple frame context. A good example is the population empirical likelihood approach (POEL), proposed by Chen and Kim (2014). As the authors noted, in the POEL approach, a single empirical likelihood is defined for the finite population. The sampling design can be incorporated into the constraint in the optimization of the POEL. Furthermore, because a single empirical likelihood is defined for the finite population, it naturally incorporates auxiliary information obtained from multiple surveys. They proved through simulation studies that the POEL estimator they propose works better than the pseudo empirical likelihood estimator for a single frame proposed by Wu (2004). Therefore, it would be interesting to consider the POEL approach to define estimators in a multiple frame context and check if they improve the results provided by the existing likelihood estimators (mainly the pseudo maximum likelihood estimator and the pseudo empirical likelihood

estimators).

- Interviewed individuals usually do not respond to part or all the items of the survey, which leads to partial or total nonresponse. If not addressed properly, nonrespose generates important biases, so results computed may be incorrect. In the multiple frame context, effects of nonresponse errors and alternatives to overcome them have been barely studied. Lohr (2007) briefly discusses the errors that may arise when conducting a multiple frame survey, including nonresponse errors. Lepkowski *et al.* (2008) focus on the nonsampling errors in dual frame surveys when one of the two frames involves telephone number. On the other hand, Lohr and Brick (2014) study the problem of the allocation in dual frame phone surveys in the presence of nonresponse. Despite these papers, literature about nonresponse in multiple frame surveys is still sparse and there is no a general approach to solve that problem. Along the development of most of the papers composing this thesis, it is assumed a full response from the interviewed individuals, so nonresponse is not a real problem. The nonresponse issue is only addressed in appendix 3, where it is considered as an additional category when analyzing the data. Therefore, nonresponse is currently being studied in a multiple frame context and a general approach to minimize its effects is investigated.

- Sometimes it is interesting obtain estimations for subgroups of the population which fulfill a specific condition. As a simple example, an investigator may be interested in compute and compare the responses to a determined question of men and women. In that situations, point estimates are easy to compute but the estimation of the variance of that estimates is not so straightforward. Problem is even more complicated in a multiple frame context, since selecting a subset of individuals of the sample s implies the reduction of the sample size n. This, in turn, could lead to small numbers of individuals of the target subpopulation in some domains, making difficult the estimation. It is clear that this issue grows with the number of domains (or, equivalently, with the number of frames) and it requires further research. Right now, techniques for the estimation in subdomains are under study.

- Development of the software should go hand to hand with theoretical advances to make feasible the resolution of practical problems. This requirement is specially important in sampling topics so that theoretical finds are available to be used in practice to achieve better results. Package *Frames2* was created for this purpose. Although initial version of the software only included the

customary dual frame estimators it has been recently updated with the multinomial estimators described in the appendix 3. Nevertheless, further updates are planned for *Frames2* to incorporate estimators for ordinal variables and the results of the current and immediate future research. On the other hand, and as it has been noted, multiple frame surveys with 3 or more sampling frames are being considered increasingly for public and private institutions when designing surveys. Typically surveys considering 3 frames composed of landline, cell and internet users, respectively, are used. But the rapid expansion of the Internet around the world is leading to 3 frame surveys composed of different list of internet users drawn from different sources. Whatever the case, 3 frame surveys are becoming a reality and so, a software similar to *Frames2* for the analysis of data coming from this type of surveys is needed. For this reason, we are working on a new R package for the estimation in a 3 frame context.

Part II

Appendices

Appendix A1

Review of estimation methods for landline and cell-phone surveys

Arcos, A., Rueda, M., Trujillo, M. and Molina, D. (2014)
Review of estimation methods for landline and cell-phone surveys.
Sociological Methods & Research.
DOI: 10.1177/0049124114546904

Abstract

The rapid proliferation of cell-phone use and the accompanying decline in landline service in recent years have resulted in substantial potential for coverage bias in landline random-digit-dial telephone surveys, which has led to the implementation of dual-frame designs that incorporate both landline and cell-phone samples. Consequently, researchers have developed methods to allocate samples and combine the data from the two frames. In this paper we review point and interval estimation methods of proportions that can be used to analyze overlapping dual frame surveys. We use data from the survey of attitudes towards immigrants and immigration (OPIA survey), a dual-frame telephone survey conducted in Andalusia, Spain, to explore these different statistical adjustments for combining landline and cell-phone samples. Our application obtains good results for calibration, fixed weight, pseudo-empirical-likelihood and single frame procedures. We recommend that one of these internally consistent estimators be used in practice.

The results of these methods of estimation show that the negative image towards immigration continues to spread.

A1.1 Introduction

Traditionally, surveys have been carried out using three main methods of data collection: face-to-face interviews, mail surveys and telephone interviews. Over the last 20 years, the picture has changed sharply. Telephone surveys have become a popular mode of data collection, especially following the creation and development of computer-assisted telephone interviewing (CATI) systems. Telephone interviews are often considered a less costly alternative to mail and face-to-face interviews and the population coverage reaches acceptable levels.

From 2000 to the present,there has been a steady increase in the use of telephone surveys, which have replaced all other data collection methods (the majority of which were face-to-face interviews). The telephone survey presents numerous advantages compared to a face-to-face one. In some subject areas (e.g. electoral studies) face-to-face surveys have been completely ousted by telephone interviewing. Moreover, studies have reported improved results from phone surveys compared with face-to-face interviews (Abascal *et al.*, 2012, Díaz de Rada, 2011).

However, telephone surveys also present some drawbacks with regard to coverage, due to the absence of a telephone in some households and the generalised use of mobile phones, which are sometimes replacing fixed (land) lines entirely (see Trujillo *et al.*, 2005, Vicente *et al.*, 2009 and Pasadas *et al.*, 2011). The potential for coverage error as a result of the exponential growth of the cell-phone-only population has led to the development of dual-frame surveys. In these designs, a traditional sample from the landline frame is supplemented with an independent sample from the banks of numbers designated for cell-phones.

By drawing samples from both cell phones and landline phones instead of from a single frame, it is possible to reduce survey costs, improve the coverage of the overall sample (Brick *et al*, 2006; Busse and Fuchs, 2012; Lu *et al.*, 2013), and potentially even increase response rates, depending on the specific survey being conducted (Opsomer, 2011).

Some surveys have used a screening dual frame survey design, in which people belonging to the landline telephone frame are removed from the cell-phone frame before sampling commences, and only people living in cell-phone-only households are interviewed (Brick *et al.*, 2007). No new statistical methods

are required to estimate totals in such a survey, since essentially a stratified sample is taken.

The screening approach can introduce a potential for bias due to nonsampling errors (Kennedy, 2007), and in many cases it may not be possible or practical to remove list-frame units from the landline frame before sampling (it is not known beforehand whether a household member sampled using one frame also belongs to the other one).

Instead, in an overlapping dual-frame survey, independent probability samples are taken from frame A (the landline frame) and frame B (the cell-phone frame). Information from the samples must be combined to estimate population quantities, and there are many options for estimators. The estimation of a population total for dual frame surveys was first investigated by Hartley (1962, 1974). Lund (1968) and Fuller and Burmeister (1972) subsequently improved on Hartley's results, and Bankier (1986) and Skinner (1991) have proposed alternative estimation techniques. More recently, Skinner and Rao (1996), Lohr and Rao (2006), Mecatti (2007), Rao and Wu (2010), Singh and Mecatti (2011) and Ranalli et al. (2013) have considered new multiple frame estimators for the population total. These methods are usually formulated under an ideal dual-frame survey setup (two frames can cover the entire target population).

In the analysis of a social survey, the response variables encountered are often discrete. For example, this would be the case for public opinion research, marketing research and government survey research. In these cases, the estimation of a proportion is a commonly used statistic for summarizing data (the proportion of voters in favour of a presidential candidate, the unemployment rate, etc.) The customary sample proportion is calculated as the percentage of individuals with a specific attribute divided by the total number of individuals in the sample. At the time of data collection, the sizes of the two frames are known. However, these two frames, in conjunction, do not usually cover the entire population, as many people do not belong to either of them. If the population size is unknown and must be estimated, the estimation for proportions is more complex than that for a total, and yet this problem has hardly been discussed in the literature on multiple frames. In this paper, we estimate the size of the conjunction of two frames and the proportion of interest in the population, using the methods described in Section 3.

After describing the OPIA survey in the second section, in the third section we consider the problem of the estimation of a proportion in our dual-frame telephone survey and then examine the effect of various estimation strategies designed to reduce the sampling error. In the fourth section we present a jackknife technique variance estimation for all estimators considered. The fifth section presents the results of the different estimation strategies in our survey dataset. Finally in the sixth section we conclude with some

thoughts about methods that could be used in future surveys that sample both landline and cell-phone numbers.

A1.2 Survey of Opinions and Attitudes of the Andalusian Population regarding Immigration (OPIA) 2013

The 2013 survey of Opinions and attitudes of the Andalusian population regarding immigration (OPIA) is a population-based survey conducted by the IESA, a public scientific research institute specialising in the social sciences. Its aim is to reflect the opinions of the Andalusian population with regard to various aspects of immigration and refugee policies in Spain and towards immigrants as a group. This survey was conducted in a period characterised by one of the most severe economic crises in the modern history of Andalusia, which has dramatically increased rates of unemployment, a situation that has notably changed attitudes towards immigration in Andalusia. This survey is based on a sample of persons drawn from both landline and cell-phone frames.

A1.2.1 Population coverage through landlines and cell phones in Andalusia

In Andalusia, the proportion of survey subjects only reachable by landline communication has decreased to below 10%. In economic good times, and due to rising numbers of internet connections, the proportion of people only reachable by cell phone also declined. However, in recent years this proportion has risen to around 20%. The number of people not reachable by phone now only represent a residual percentage of the population (less than 2%).

Table A1.1: Coverage in 2013. Source: Survey of Information Technologies in Households (INE).

Both	69.4%
Cell only	9.6%
Land only	19.7%
No phone	1.3%

The distributions of landlines and cell phones vary considerably depending on the age of the population. Figure 2 shows that, taking into account only people for whom the availability of a landline depends on their own decision, that is, not considering people living with their parents, the younger the

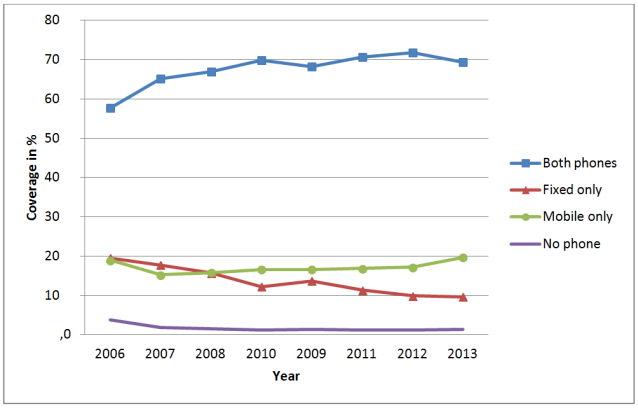


Figure A1.1: Evolution of landline and cell phone coverage for people over 16 years old.
Source: Survey on the Equipment and Use of Information and Communication Technologies (ICT - H) in Households.

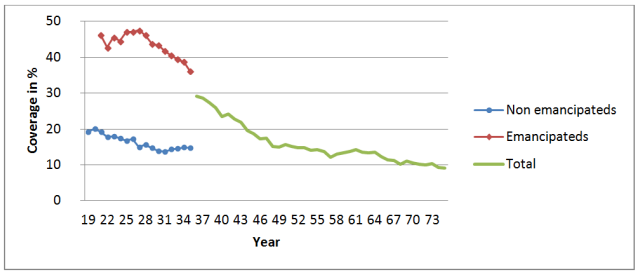
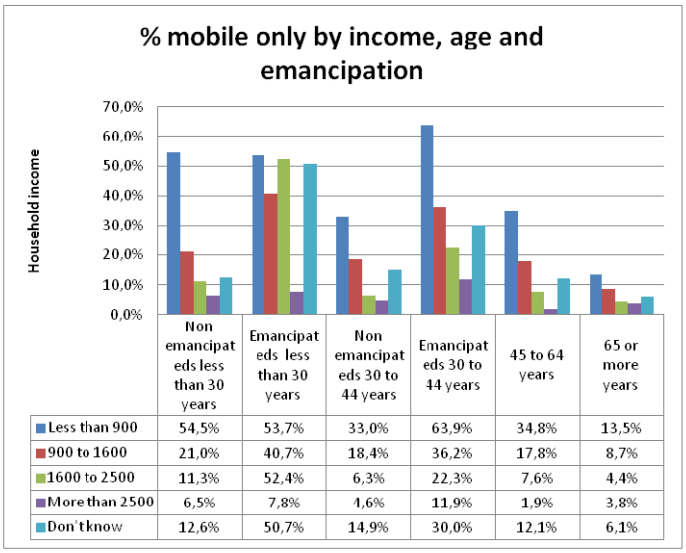


Figure A1.2: Percentage of people with only cell phone, by age.
Source: Survey on the Equipment and Use of Information and Communication Technologies (ICT - H) in Households.

population, the higher the percentage having only a cell phone. This value exceeds 40% for people aged under 33 years.

A worrying issue in this respect, due to the difficulties posed in correcting it, is the income gap between those with only a cell phone and the rest of the population (Vicente and Reis, 2009). In Figure 3 it can be seen, taking into account the age and the state of emancipation, that there are very large differences in the percentages of people who have only a cell phone, depending on personal income. For example, for people living independently and aged between 30 and 44 years, 60% of individuals have only a cell phone when their household income is below 900 euros, and this percentage is 10% when their income exceeds 2,500 euros.

In this survey, the IESA decided to carry out telephone interviews with adults using both landlines and cell phones. Taking into account the time and budget available, 2402 interviews were performed by qualified interviewers, specially trained in survey techniques. The number of interviews to be conducted



	Non emancipateds less than 30 years	Emancipateds less than 30 years	Non emancipateds 30 to 44 years	Emancipateds 30 to 44 years	45 to 64 years	65 or more years
Less than 900	54,5%	53,7%	33,0%	63,9%	34,8%	13,5%
900 to 1600	21,0%	40,7%	18,4%	36,2%	17,8%	8,7%
1600 to 2500	11,3%	52,4%	6,3%	22,3%	7,6%	4,4%
More than 2500	6,5%	7,8%	4,6%	11,9%	1,9%	3,8%
Don't know	12,6%	50,7%	14,9%	30,0%	12,1%	6,1%

Figure A1.3: Percentage of population with only cell phone, by income, age and emancipation.
Source: Survey on the Equipment and Use of Information and Communication Technologies (ICT - H) in Households.

via landline and via cell phone was determined by calculating the optimum proportion (in the sense of minimum variance) for each type of telephone, taking into account the (Pasadas and Trujillo, 2013) costs and the percentage of possession of each type of device (following Hartley, 1962). As a result, the sample sizes ascertained were 1919 for landlines and 483 for cell phones. The interviews were carried out by the Statistics and Surveys sections of IESA from 22 April to 13 May 2013, using Computer Assisted Telephone Interviewing (CATI) data input techniques.

A1.2.2 Descriptions of frames and sampling designs

Following Hartley's classical notation (1962), two samples are drawn independently from two frames, A and B. Let $a = A \cap \overline{B}$, $b = \overline{A} \cap B$, and $ab = A \cap B$, where $\overline{(\cdot)}$ denotes the complement of a set. From frame A, land-phone, a stratified sample s_A of size n_A was drawn. Probability-based random-digit-dial (RDD) telephone survey is performed in frame B, cell-phone, and a sample s_B of size n_B is drawn using a simple random sampling without replacement design, SRSWOR.

Sample sizes of land (A) and cell (B) phones are $n_A = 1919$ and $n_B = 483$. Domain sample sizes are: in the overlapping population $n_{ab} = 1727$ for the sample $s_{ab} = (s_A \cap ab)$, $n_{ba} = 237$ for the sample $s_{ba} = (s_B \cap ab)$ and $n_b = 246$ for the cell phone sample $s_b = s_B \cap b$ and $n_a = 192$ for the

land phone sample $s_a = s_A \cap a$. The total sample is $s = s_A \cup s_B = s_a \cup s_{ab} \cup s_{ba} \cup s_b$, and its size is $n = n_A + n_B = n_a + n_{ab} + n_{ba} + n_b = 2402$.

At the time of data collection, frame sizes of land (A) and cell (B) phones were $N_A = 4982920$ and $N_B = 5707655$ and the total population size was $N = 6350916$. The domain population sizes were: $N_{ab} = 4339659$ for the overlap domain, $N_a = 643261$ for land phones and $N_b = 1367996$ for cell phones. (source ICT-H 2012, Survey on the Equipment and Use of Information and Communication Technologies in Households, INE, National Statistical Institute, Spain.)

Table A1.2: Sample sizes for the OPIA survey. Land and Cell in the columns refer to the frame from which the units were chosen, while in the rows, they refer to frame in which the units actually reside.

	Land	Cell	Total
Both	1727	237	1964
Cell		246	246
Land	192		192
Total	1919	483	2402

The land-phone sample was also stratified by provinces in the region of Andalusia, as shown in Table A1.3.

Table A1.3: Stratification in land-phone sample

Province	Almería	Cádiz	Córdoba	Granada	Huelva	Jaén	Málaga	Sevilla
N_h^A(*)	353787	767370	508258	558087	308941	423548	872011	1190918
n_h^A	262	210	252	256	275	263	207	194

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

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.

A1.2.3 Initial weighting adjustments

This section describes the procedures used to create the weights for each sample. The base weights are the ratio of the number of telephone numbers in the frame to the number sampled. The weights were further adjusted to account for people who had multiple chances of being sampled because they had more

than one telephone number.

First order inclusion probabilities were computed from a stratified random design in frame A and modified taking into account the number of fixed lines (L_{hk}) and adults in the household (A_{hk}) as follows: $\pi^A_{hk} = \frac{n^A_h L_{hk}}{N^A_h A_{hk}}$. The design weights were computed as $d^A_{hk} = 1/\pi^A_{hk}$ for all h and k. A simple random sample without replacement, SRSWOR, was drawn from frame B and first order inclusion probabilities were computed and modified given the number of cell-phone numbers per individual (M_k) as $\pi^B_k = \frac{n_B M_k}{N_B}$, for all k. The design weights were computed as $d^B_k = 1/\pi^B_k$.

A1.3 Estimation in dual frame telephone surveys

We consider the problem of estimating the population proportion $P = N^{-1}\sum_{k=1}^{N} y_k$, where y_k is an attribute indicator for unit k, i.e., $y_k = 1$ if unit k has the attribute of interest, and $y_k = 0$ otherwise. The number of population units belonging to the group of interest is denoted by $Y = \sum_{k=1}^{N} y_k$.

If the population size is known, an estimator $\widehat{P}$ of the population can easily be obtained from the total estimator $\widehat{Y}$ as the ratio $\widehat{P} = \widehat{Y}/N$. In cases where the population size is unknown, $\widehat{P} = \widehat{Y}/\hat{N}$ is an estimator of P, where $\hat{N}$ is an estimate of the population size N (this situation can arise in practice when, for example, the sampling frames available do not cover the entire target population).

We now present an overview of the estimation procedures of $\widehat{Y}$ used in this survey.

A1.3.1 Single-frame approach

Bankier (1986) and Kalton and Anderson (1986) proposed estimators that treated all the observations as if they had been sampled from a single frame, with adjusted weights in the intersection domain relying on the inclusion probabilities for each frame. In those situations, as in our example, in which we know the inclusion probability of the units in the sample under both sampling designs, the weights are defined as follows for all units in frame A and in frame B:

$$d^{sf}_k = \begin{cases} d^A_k & \text{if} \quad k \in a \\ (1/d^A_k + 1/d^B_k)^{-1} & \text{if} \quad k \in ab \\ d^B_k & \text{if} \quad k \in b \end{cases} \tag{A1.1}$$

Note that the units in the overlap domain, which are expected to be selected with a probability

$(\pi_k^A + \pi_k^B)$, have equal weights in frame A and in frame B.

Single frame estimator (SF). Kalton and Anderson's (1986) single frame estimator is:

$$\hat{Y}^{SF} = \sum_{k=1}^{n} d_k^{sf} y_k. \tag{A1.2}$$

The single frame weights are the same for all response variables, and so the estimators are internally consistent. For complex surveys, however, single frame estimators may not be efficient. Skinner (1991) provides a theoretical study of the efficiency of the raking ratio estimator for multiple-frame survey. For the calculation of an unbiased estimator of the variance of a single-frame estimator, we adopted the approach proposed by Rao and Skinner (1996)

$$\hat{V}(\hat{Y}^{SF}) = \hat{V}(\tilde{z}_k^A) + \hat{V}(\tilde{z}_k^B), \tag{A1.3}$$

where $\tilde{z}_k^A = \delta_k(a) y_k + (1 - \delta_k(a)) y_k \frac{\pi_k^A}{\pi_k^A + \pi_k^B}$, $\tilde{z}_k^B = \delta_k(b) y_k + (1 - \delta_k(b)) y_k \frac{\pi_k^B}{\pi_k^A + \pi_k^B}$ and $\hat{V}(\cdot)$ denotes the Horvitz-Thompson variance estimator (see Särndal et al., 1992) with $\delta_k(a) = 1$ if $k \in a$ and 0 otherwise, $\delta_k(ab) = 1$ if $k \in ab$ and 0 otherwise, $\delta_k(ba) = 1$ if $k \in ba$ and 0 otherwise and $\delta_k(b) = 1$ if $k \in b$ and 0 otherwise.

Calibration estimator (CAL). In the OPIA survey, N_A, N_B and N_{ab} are all known. We can define a calibration estimator on (N_a, N_{ab}, N_b):

$$\hat{Y}^{CAL} = \sum_{k=1}^{n} w_k^{cal} y_k \tag{A1.4}$$

with weights w^{cal} verified to be close to the design weights d_k^{sf} and that reproduce the known totals (N_a, N_{ab}, N_b), that is, $\hat{N}_a^{CAL} = \sum_{k=1}^{n} w_k^{cal} \delta_k(a) = N_a$, $\hat{N}_b^{CAL} = \sum_{k=1}^{n} w_k^{cal} \delta_k(b) = N_b$ and $\hat{N}_{ab}^{CAL} = \sum_{k=1}^{n} w_k^{cal} \delta_k(ab) = N_{ab}$. All the distance measures taken to define "closeness" provide the same set of calibration weights, because the minimization problem has an analytic solution irrespective of the distance function employed (see Ranalli *et al.* 2013 for details).

An estimator of the variance of calibration estimator can be obtained using the residuals of regression

of y on $\mathbf{x} = (\delta_k(a), \delta_k(ab), \delta_k(b))$ as the y-variable in expression (A1.24).

Single Frame Raking Ratio (SFRR). The single-frame estimator (SF) does not use any auxiliary information about the population totals N_A and N_B, but can be adjusted through any of the raking ratio estimations. Skinner (1991) and Rao and Skinner (1996) showed that the raking procedures in fact converge to give the explicit estimator

$$\hat{Y}^{SFRR} = \frac{N_A - \hat{N}_{ab}^{RR}}{\hat{N}_a^{SF}} \hat{Y}_a^{SF} + \frac{\hat{N}_{ab}^{RR}}{\hat{N}_{ab}^{SF}} \hat{Y}_{ab}^{SF} + \frac{N_B - \hat{N}_{ab}^{RR}}{\hat{N}_b^{SF}} \hat{Y}_b^{SF} \tag{A1.5}$$

where $\hat{N}_{ab}^{RR}$ is the smallest root of the quadratic equation $\hat{N}_{ab}^{SF} x^2 - \Big[\hat{N}_{ab}^{SF}$ $(N_A + N_B) + \hat{N}_a^{SF} \hat{N}_b^{SF}\Big] x + \hat{N}_{ab}^{SF} N_A N_B = 0$.

If N_{ab} is not known, a calibration estimator can be defined on (N_A, N_B):

$$\tilde{Y}^{CAL} = \sum_{k=1}^{n} \tilde{w}_k^{cal} y_k \tag{A1.6}$$

with weights $\tilde{w}^{cal}$ verified to be close to the design weights d_k^{sf} and that reproduce the known totals (N_A, N_B), that is, $\tilde{N}_A^{CAL} = \sum_{k=1}^{n} \tilde{w}_k^{cal} \delta_k(A) = N_A$ and $\tilde{N}_B^{CAL} = \sum_{k=1}^{n} \tilde{w}_k^{cal} \delta_k(B) = N_B$. This estimator is the same as SFRR in (A1.5) if the "raking" method is used in calibration.

The variance for the single frame calibration estimator is then determined using the residuals of regression of y on $\mathbf{x} = (\delta_k(A), \delta_k(B))$ as the y-variable in expression (A1.24).

A1.3.2 Dual-frame approach

In situations in which we do not know the inclusion probability of the units in the sample under both sampling designs, dual-frame methods can be considered. For comparison, these methods are also considered in our example.

We can write

$$Y = Y_a + \eta Y_{ab} + (1 - \eta) Y_{ba} + Y_b, \tag{A1.7}$$

where $Y_a = \sum_{j \in a} y_j$, $Y_{ab} = \sum_{j \in ab} y_j$, $Y_{ba} = \sum_{j \in ba} y_j$ and $Y_b = \sum_{j \in b} y_j$.

Fixed weight adjustment (FWA). The simplest weight modification to preserve approximate unbiasedness, as described by Hartley (1962), yields

$$\hat{Y}(\theta) = \hat{Y}_a + \theta\hat{Y}_{ab} + (1-\theta)\hat{Y}_{ba} + \hat{Y}_b \tag{A1.8}$$

Brick *et al.* (2006) used $\theta = 1/2$ in their study of a dual-frame survey in which frame A was a landline telephone frame and frame B was a cell-phone frame. For this purpose, the value of $\theta = 1/2$ is frequently recommended (see, for example, Mecatti 2007). This estimator is denoted by

$$\hat{Y}^{FWA} = \hat{Y}_a + (1/2)\hat{Y}_{ab} + (1/2)\hat{Y}_{ba} + \hat{Y}_b \tag{A1.9}$$

In order to calculate an estimator of the variance, we have taken into account that samples from frames A and B are drawn independently and that the value for θ is fixed. Thus,

$$\hat{V}(\hat{Y}(\theta)) = \hat{V}(\hat{Y}_a + \theta\hat{Y}_{ab}) + \hat{V}((1-\theta)\hat{Y}_{ba} + \hat{Y}_b) \tag{A1.10}$$

where (A1.24) is used to compute the variance estimations.

Hartley (HAR) (1962, 1974) proposed choosing θ in (A1.8) so that the variance of $\hat{Y}(\theta)$ would be minimized. The optimizing value of θ is

$$\theta_{opt} = \frac{V(\hat{Y}_{ba}) + cov(\hat{Y}_b, \hat{Y}_{ba}) - cov(\hat{Y}_a, \hat{Y}_{ab})}{V(\hat{Y}_{ab}) + V(\hat{Y}_{ba})} \tag{A1.11}$$

and the estimator has the form

$$\hat{Y}^{HAR}(\theta_{opt}) = \hat{Y}_a + \theta_{opt}\hat{Y}_{ab} + (1-\theta_{opt})\hat{Y}_{ba} + \hat{Y}_b \tag{A1.12}$$

However, this optimal estimator is a function of the variances and covariances of the estimated domain totals and then the optimal estimates will differ for different response variables.

In cases where estimation of θ_{opt} is outside $[0, 1]$, approximation

$$\theta_{opt} \simeq \frac{V(\hat{Y}_{ba})}{V(\hat{Y}_{ab}) + V(\hat{Y}_{ba})} \tag{A1.13}$$

can be used instead. In our example, using (A1.24) to estimate the three variances found in the latter expression of θ_{opt}, we can obtain an estimation for the θ_{opt} without using second-order inclusion probabilities. The variance estimator for the Hartley estimator can be obtained by replacing θ in (A1.10) for the θ_{opt} value given in (A1.11).

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

$$\hat{Y}^{FB}(\beta) = \hat{Y}_a + \beta_1 \hat{Y}_{ab} + (1-\beta_1)\hat{Y}_{ba} + \hat{Y}_b + \beta_2(\hat{N}_{ab} - \hat{N}_{ba}) \tag{A1.14}$$

where β_1 and β_2 are selected to minimize $V(\hat{Y}_{FB}(\beta))$. In this case, and as with Hartley's estimator, a new set of weights must be calculated for each response variable, leading to the inconsistency of the estimator. Optimum values depend on covariances among the Horvitz-Thompson estimators and it is also possible to obtain values of β_1 outside $[0, 1]$. Moreover, it is not possible to estimate the population size N using the FB estimator, because the minimization process requires the inversion of a singular matrix.

Pseudo-Maximum Likelihood (PML). Skinner and Rao (1996) proposed modifying the maximum likelihood estimator for a simple random sample suggested by Fuller and Burmeister (1972) to obtain a pseudo-maximum likelihood (PML) estimator for a complex design. The PML estimator, unlike the Hartley and Fuller-Burmeister estimators, is linear in y and is of the form

$$\hat{Y}^{PML}(\theta) = \frac{N_A - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a}\hat{Y}_a + \frac{\hat{N}_{ab}^{PML}(\theta)}{\hat{N}_{ab}(\theta)}\hat{Y}_{ab}(\theta) + \frac{N_B - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b}\hat{Y}_b \tag{A1.15}$$

where $\hat{Y}_{ab}(\theta) = \theta\hat{Y}_{ab} + (1-\theta)\hat{Y}_{ba}$, $\hat{N}_{ab}(\theta) = \theta\hat{N}_{ab} + (1-\theta)\hat{N}_{ba}$ and $\hat{N}_{ab}^{PML}(\theta)$ is the smallest root of the quadratic equation $[\theta/N_B + (1-\theta)/N_A]\, x^2 - \left[1 + \theta\hat{N}_{ab}/N_B + (1-\theta)\hat{N}_{ba}/N_A\right] x + \hat{N}_{ab} = 0$. Skinner and Rao (1996) suggested choosing θ to minimize the asymptotic variance of $\hat{N}_{ab}^{PML}(\theta)$, with

$$\hat{\theta} = \frac{N_a N_B \widehat{V}(\widehat{N}_{ba})}{N_a N_B \widehat{V}(\widehat{N}_{ba}) + N_b N_A \widehat{V}(\widehat{N}_{ab})} \tag{A1.16}$$

or estimate it as

$$\theta \simeq \frac{V(\hat{N}_{ba})}{V(\hat{N}_{ab}) + V(\hat{N}_{ba})} \tag{A1.17}$$

In practice the variances in (A1.16) are unknown and must be estimated from the data. The PML estimator uses the same set of weights for each response variable and thus avoids some of the difficulties associated with the Hartley and Fuller-Burmeister estimators.

To estimate the variance of the PML estimator, we followed the method proposed by Rao and Skinner (1996), which provides a consistent estimator of variance in the form

$$\hat{V}(\hat{Y}^{PML}) = \hat{V}(\tilde{z}_k^A) + \hat{V}(\tilde{z}_k^B), \tag{A1.18}$$

where, in this case, $\tilde{z}_k^A = y_k - \frac{\hat{Y}_a}{\hat{N}_a}$ if $k \in s_a$ and $\tilde{z}_k^A = \theta\left(y_k - \frac{\hat{Y}_{ab}}{\hat{N}_{ab}}\right) + \hat{\lambda}\hat{\phi}$ if $k \in s_{ab}$, where θ is calculated according to (A1.16), $\hat{\lambda} = \frac{\hat{Y}_{ab}}{\hat{N}_{ab}} - \frac{\hat{Y}_a}{\hat{N}_a} - \frac{\hat{Y}_b}{\hat{N}_b}$ and $\hat{\phi} = \frac{n_A\hat{N}_b}{n_A\hat{N}_b + n_B\hat{N}_a}$. Similarly, we can define $\tilde{z}_k^B = y_k - \frac{\hat{Y}_b}{\hat{N}_b}$ if $k \in s_b$ and $\tilde{z}_k^B = (1-\theta)\left(y_k - \frac{\hat{Y}_{ba}}{\hat{N}_{ab}}\right) + \hat{\lambda}(1-\hat{\phi})$ if $k \in s_{ba}$.

Pseudo-Empirical-Likelihood (PEL). 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 and considering an estimation of the population mean of y,

$$\hat{\bar{Y}}^{PEL}(\theta) = (N_a/N)\hat{\bar{Y}}_a + (\theta)(N_{ab}/N)\hat{\bar{Y}}_{ab} + (N_{ab}/N)(1-\theta)\hat{\bar{Y}}_{ba} + (N_b/N)\hat{\bar{Y}}_b, \tag{A1.19}$$

where $\theta \in (0,1)$ is a fixed constant to be specified and $\hat{\bar{Y}}_a = \sum_{k\in s_a}\hat{p}_{ak}y_k$, $\hat{\bar{Y}}_b = \sum_{k\in s_b}\hat{p}_{bk}y_k$ and $\hat{\bar{Y}}_{ab} = \sum_{k\in s_{ab}}\hat{p}_{abk}y_k = \hat{\bar{Y}}_{ba}$. The weights maximize the pseudo empirical likelihood and verify $\sum_{k\in s_a} p_{ak} = 1$, $\sum_{k\in s_{ab}} p_{abk} = 1$, $\sum_{k\in s_{ba}} p_{bak} = 1$, $\sum_{k\in s_b} p_{bk} = 1$, and the additional constraint induced by the common domain mean $\bar{Y}_{ab} = \bar{Y}_{ba}$. In this case, we use the same estimation for θ as the one proposed in (A1.17).

Instead of calculating the explicit variance of the estimator, confidence intervals are obtained using the bi-section method described by Wu (2005). This method constructs intervals in the form $\theta | r_{ns}(\theta) < \chi_1^2(\alpha)$, where $\chi_1^2(\alpha)$ is the $1-\alpha$ quantile from a χ^2 distribution with one degree of freedom and r_{ns} represents the pseudo empirical log likelihood ratio statistic, which can be obtained as the difference of two PEL functions.

A1.4 Jackknife variance estimation

We also use jackknife estimation to determine the variance of the estimators compared (Wolter, 2007). The variance estimators presented in the third section can be computed in many different ways, depending on the specific estimator. Moreover, in small samples, they may poorly estimate the variability of estimators because they estimate the asymptotic variance rather than the exact variance. Instead, the jackknife approach is a common method for variance estimation that can be used whatever the estimator. Thus, estimated variances obtained through this method can be used to compare the efficiencies of the estimators. For the sake of brevity, in this section dual or single-frame estimators are denoted by $\hat{Y}_c$.

In the case of a stratified design, as in frame A, let frame A be divided into H strata and let stratum h have N_{Ah} observation units of which n_{Ah} are sampled. Then, a jackknife variance estimator of $\hat{Y}_c$ with an approximate finite-population correction is given by

$$V_J^A(\hat{Y}_c) = \sum_{h=1}^{H} \left(1 - \frac{n_{Ah}}{N_{Ah}}\right) \frac{n_{Ah}-1}{n_{Ah}} \sum_{i \in s_{Ah}} (\hat{Y}_c^A(hi) - \overline{Y}_c^{Ah})^2 \tag{A1.20}$$

where $\hat{Y}_c^A(hi)$ is the value taken by estimator $\hat{Y}_c$ after dropping unit i of stratum h from sample s_{Ah}, $\overline{Y}_c^{Ah}$ is the average of these n_{Ah} values.

If we consider a non stratified design, as in frame B, the jackknife estimator for the variance of $\hat{Y}_c$ with an approximate finite-population correction may be given by

$$V_J^B(\hat{Y}_c) = \frac{n_B - 1}{n_B}(1 - \frac{n_B}{N_B}) \sum_{i \in s_B} (\hat{Y}_c^B(i) - \overline{Y}_c^B)^2 \tag{A1.21}$$

where $\hat{Y}_c^B(i)$ is the value taken by estimator $\hat{Y}_c$ after dropping unit i from s_B and $\overline{Y}_c^B$ is the average of $\hat{Y}_c^B(i)$ values (see Wolter, 2007).

For any estimator $\hat{Y}_c$ in the single or dual frame approach, we compute $\hat{Y}_c(i), i = 1, \ldots, n$. Then, the pseudo values $\hat{Y}_c(i)$ are separated into those from frame A and from frame B and V_J^A and V_J^B are computed. Finally, due to the independence, $V_J = V_J^A + V_J^B$.

A1.5 Results for the OPIA Survey

To examine the performance of the dual-frame estimation methods in practice, we applied them to the dataset from the OPIA survey.

Three main variables are included in this study, related to "goodness of immigration", "amount of immigration" and "confidence in immigration". The variables are the answers to the following questions:

- *And in relation to the number of immigrants currently living in Andalusia, do you think there are ...?: Too many, A reasonable number, Too few*
- *In general, do you think that for Andalusia, immigration is ...?: Very bad, Bad, Neither good nor bad, Good, Very good*
- *In general, how much confidence do you have in immigrants? None at all, Very little, It depends, Quite a lot, Very much*

Each category of each variable is treated separately as an attribute so that for any of the attributes of interest, $y_{kI} = 1$ if the k-th individual presents the attribute I and $y_{kI} = 0$ otherwise. The proportions for all the main variables are computed using $\hat{P}_I = \frac{\hat{Y}_I}{\hat{N}}$, where $\hat{Y}_I$ is the estimated total of units in the population with the attribute of interest I and $\hat{N}$ is an estimate of the population size N. For example, using the single frame estimator (A3.9) we estimate the population total and the population size as:

$$\hat{Y}_I^{SF} = \sum_{k=1}^{n} d_k^{sf} y_{kI} \quad \text{and} \qquad \hat{N}^{SF} = \sum_{k=1}^{n} d_k^{sf}, \tag{A1.22}$$

respectively, and similarly for the other estimators. For the FB estimator, the matrix to solve the minimum variance is singular in estimating the population size N and this estimator is not included.

The weights w_k^{cal} of the calibration estimator (A1.4) verify that

$$\sum_{k=1}^{n} w_k^{cal}\delta_k(a) = 643261, \quad \sum_{k=1}^{n} w_k^{cal}\delta_k(ab) = 1367996, \quad \sum_{k=1}^{n} w_k^{cal}\delta_k(b) = 4339659. \tag{A1.23}$$

As Särndal (2007) says, the calibration gives a unique weighting system, one that is perfectly clear and transparent, and applicable to all study variables.

In the dual-frame approach, there is no single $\hat{\theta}_{opt}$ for the HAR estimator, since it depends on the values of each study variable. For the PEL estimator, the value for $\hat{\theta}$ in (A1.17)(applicable to all study

variables) is $\hat{\theta} = 0.729684$, whereas with the FWA estimator we use $\theta = 1/2$. For the PML estimator, the value for θ in (A1.16) is $\hat{\theta} = 0.620662$.

All dual-frame estimators have one thing in common: the weighting of the estimations for the overlap domain, either with 1/2 or with one of estimations of θ in (A1.11), (A1.13) or (A1.16). In single-frame estimators, the weighting is given by probabilities under both sampling designs.

All the estimators considered in this paper require estimates of the domain sizes N_a, N_b and N_{ab}. The estimates for the sizes of the population domains are obtained using the Horvitz-Thompson estimator. For domain a, the population size N_a is estimated by $\widehat{N}_a = \sum_{k=1}^{n_A} d_k^A \delta_k(a)$ where $\delta_k(a) = 1$ if $k \in a$ and 0 otherwise. For domain ab, there are two options: a) the population size N_{ab} is estimated by $\widehat{N}_{ab} = \sum_{k=1}^{n_A} d_k^A \delta_k(ab)$ where $\delta_k(ab) = 1$ if $k \in ab$ and 0 otherwise and b) the population size N_{ab} is estimated by $\widehat{N}_{ba} = \sum_{k=1}^{n_B} d_k^B \delta_k(ab)$. For domain b, the population size N_b is estimated by $\widehat{N}_b = \sum_{k=1}^{n_B} d_k^B \delta_k(b)$ where $\delta_k(b) = 1$ if $k \in b$ and 0 otherwise.

In a similar way, we denote the Horvitz-Thompson estimator of any y variable in domain a as $\widehat{Y}_a = \sum_{k=1}^{n_A} d_k^A \delta_k(a) y_k$ and similarly for the others. In the present survey the following results are obtained:

Table A1.4: Estimates of domain sizes and coefficients of variation

Domain	Estimate	CV
a	493776	0,084
ab	4646468	0,020
ba	3117703	0,049
b	3227202	0,047

The variances in Table A1.4 are computed using Deville's method (Deville, 1993) to avoid second-order probabilities (although in this case it is possible to easily compute them). This method yields, given a y-variable whose population total Y is estimated using the Horvitz-Thompson estimator based on a sample s, $\hat{Y} = \sum_s y_k/\pi_k$, the following variance estimator:

$$\hat{V}(\hat{Y}) = \frac{1}{1 - \sum_{k \in s} a_k^2} \sum_{k \in s} (1 - \pi_k) \left(\frac{y_k}{\pi_k} - \sum_{l \in s} a_l \frac{y_l}{\pi_l} \right)^2 \tag{A1.24}$$

where $a_k = (1 - \pi_k) / \sum_{l \in s} (1 - \pi_l)$.

Tables A1.5, A1.6 and A1.7 show the point and 95% confidence level estimation of proportions of the main variables. Two different sets of confidence intervals are calculated: one, based on the jackknife

variance estimation described in the fourth section and the other, based on the variance estimations described in the third section. Other tables could be obtained if finite population correction factors were used in jackknife variance estimation, but they are not included here because the results would be very similar.

For the outcomes shown in Tables A1.5, A1.6 and A1.7, we obtained the following findings:

- There are no important differences between the estimates produced with the single or dual frame approach.

- Among all the estimation strategies, the calibration method performs best, and produces the smallest confidence interval. Calibration estimation can be implemented easily using existing software for single-frame populations. There are several R packages for obtaining estimations using the calibration technique, as the *sampling* package.

- The jackknife method often produces better intervals than methods based on the estimated variance given by the authors (except for the PEL intervals)

At the time of data collection, the frame sizes for land phones (A) and cell phones (B) were $N_A = 4982920$ and $N_B = 5707655$ and the overlap domain was size $N_{ab} = 4339659$. We also studied the effect on estimation of using different values for frame and overlap domain sizes extracted from different sources. For this purpose, we considered the three sets of sizes shown in Table 8. The data were obtained from the Survey on the Equipment and Use of Information and Communication Technologies in Households (conducted by the Spanish National Institute of Statistics) and from the IESA Households Survey conducted in 2012 and 2013. Using four of the estimators described in the third section, we computed the three possible estimations, the average values and the coefficients of variation. The results of this are shown in Tables 9 and 10 for the three main variables.

The estimates obtained by each method, using different values of frame sizes obtained from 3 sources, are, in general, similar. It is concluded that the estimators are only slightly influenced by the source used to estimate the population sizes for landline and cell phones.

A1.6 Conclusions

This paper addresses some of the issues involved in using dual-frame methods for landline and cell-phone surveys. Multiple frame surveys are very useful when it is not possible to guarantee complete coverage of the target population, and may result in considerable cost savings in comparison with a single-frame design with comparable precision. However, this technique is not often applied by national statistical agencies or by private survey agencies due to its complexity and the difficulties inherent in analyzing multiple-frame surveys with standard survey software.

Several estimators have been proposed and the first question to be considered is how to choose the most suitable one for this application.

Calibration, fixed weight, PML and single-frame estimators all give internal consistency, since the same set of adjusted weights is used for all variables. In our application, good results were obtained with these procedures. We recommend that an internally-consistent estimators be used. With repeated surveys, the simplicity and transparency of a fixed-weight estimator may be preferred. Fixed-weight adjustments may make year-to year comparisons easier in an annual survey, where the domain proportions are relatively constant over time. Fixed-weight estimators are also more amenable to weight adjustments for non-response and domain misclassification. Standard survey software may then be used to estimate population proportions and totals using the modified weights.

On the other hand, variance estimation is more complicated when dual-frame estimators are used. Resampling methods such as jackknife estimation may then be used to estimate variances. Jackknife intervals are easy to compute and give accurate intervals.

The dual-frame estimates obtained from the variables considered in this study suggest that the use of different values for frame and overlap domain sizes extracted from different sources had no substantial impact on the level of efficiency obtained.

In this study, the use of auxiliary variables was not considered for estimating the study variables. The use of demographic variables such as age, income or emancipation in the calibration and pseudo-empirical-likelihood methods can improve the estimates, because these variables can have a considerable impact on the distribution of landlines and cell phones.

We also highlight the need to implement these methods in both commercial and non-commercial software for survey estimation. In this respect, we are now working on an R package for point and

interval estimation for a two-frame estimator.

Finally, let us note that the results obtained in applying these methods in the OPIA survey indicate that negative views towards immigration continue to spread, and that currently 59-61% of those surveyed in Andalusia state that immigration is bad or very bad for the region (in the previous edition of the study, in 2011, the corresponding figure was 58 %, and in the first such survey, in 2005, it was only 51%). Perceptions regarding the number of immigrants, however, have changed in the opposite direction: there is now a lower percentage of people who say there are too many immigrants (from 51 % in 2011 to current levels of 40-42 %), while the other scores have risen slightly.

Table A1.5: Point and 95% confidence level estimation of proportions using several methods for variance estimation. Main variable: "goodness of immigration"

In general, do you think that for Andalusia, immigration is ...?							
		Jackknife variance			Analytical variance		
Estimator	PROP	LB	UB	LEN	LB	UB	LEN
				Very bad			
SF	13.72	11.57	15.87	4.31	11.57	15.87	4.30
SFRR	13.90	11.84	15.95	4.11	11.09	16.12	5.03
CAL	13.35	11.69	15.01	3.33	10.72	15.97	5.25
FWA	13.70	11.66	15.74	4.08	11.68	16.30	4.62
HAR	13.44	11.66	15.23	3.58	11.64	15.98	4.34
PML	13.87	11.76	15.97	4.21	11.57	16.87	5.30
PEL	13.62	11.71	15.53	3.82	12.89	15.86	2.97
				Bad			
SF	47.24	43.72	50.77	7.05	43.79	50.70	6.91
SFRR	47.39	44.43	50.35	5.92	43.98	51.14	7.16
CAL	46.92	44.52	49.33	4.81	43.18	50.66	7.48
FWA	45.48	42.24	48.72	6.49	43.10	50.16	7.06
HAR	46.16	43.19	49.12	5.93	43.06	49.93	6.87
PML	46.43	43.02	49.84	6.82	42.96	50.95	7.99
PEL	45.95	43.24	48.65	5.41	45.22	50.79	5.57
				Neither good nor bad			
SF	4.85	3.54	6.16	2.61	3.54	6.16	2.61
SFRR	4.47	3.42	5.51	2.09	3.08	6.19	3.11
CAL	4.75	3.75	5.74	1.99	3.13	6.37	3.24
FWA	4.20	3.18	5.23	2.05	3.17	5.88	2.71
HAR	4.60	3.59	5.62	2.03	3.09	5.68	2.59
PML	4.34	3.30	5.38	2.08	2.44	5.43	2.99
PEL	4.33	3.30	5.36	2.06	2.81	5.21	2.40
				Good			
SF	28.35	25.56	31.14	5.58	25.58	31.13	5.55
SFRR	28.22	25.87	30.57	4.70	25.00	31.33	6.33
CAL	28.98	26.86	31.11	4.25	25.68	32.29	6.61
FWA	30.46	27.74	33.19	5.45	25.98	31.85	5.87
HAR	29.93	27.52	32.34	4.82	25.71	31.32	5.61
PML	29.36	26.92	31.81	4.89	25.19	31.81	6.62
PEL	29.96	27.49	32.43	4.94	25.05	29.96	4.91
				Very good			
SF	2.18	1.36	3.00	1.63	1.36	3.00	1.64
SFRR	2.10	1.41	2.79	1.38	1.12	3.08	1.96
CAL	2.16	1.51	2.82	1.31	1.14	3.19	2.05
FWA	2.14	1.35	2.93	1.58	1.25	3.03	1.78
HAR	2.11	1.43	2.78	1.35	1.29	2.94	1.65
PML	2.08	1.36	2.80	1.44	0.98	3.05	2.07
PEL	2.12	1.35	2.88	1.53	1.39	2.54	1.15

Table A1.6: Point and 95% confidence level estimation of proportions using several methods for variance estimation. Main variable: "Amount of immigration"

In relation to the number of immigrants currently living in Andalusia, do you think there are ...?							
Estimator		Jackknife variance			Analytical variance		
	PROP	LB	UB	LEN	LB	UB	LEN
	Too many						
SF	42.31	38.95	45.66	6.71	38.97	45.64	6.67
SFRR	40.69	37.90	43.48	5.59	37.79	44.90	7.11
CAL	40.97	38.61	43.34	4.74	37.26	44.69	7.43
FWA	40.26	37.28	43.24	5.97	39.10	45.94	6.84
HAR	39.92	37.26	42.59	5.33	38.75	45.42	6.67
PML	40.44	37.29	43.59	6.30	38.12	45.86	7.74
PEL	41.05	38.37	43.73	5.36	39.78	43.93	4.15
	A reasonable number						
SF	45.81	42.44	49.19	6.74	42.51	49.12	6.61
SFRR	47.85	44.99	50.72	5.73	43.05	50.15	7.10
CAL	47.03	44.63	49.43	4.79	43.32	50.74	7.42
FWA	47.91	44.59	51.23	6.64	42.03	48.82	6.79
HAR	48.43	45.41	51.45	6.04	42.02	48.63	6.61
PML	47.95	44.88	51.02	6.14	41.82	49.35	7.53
PEL	46.72	44.00	49.43	5.43	43.44	47.96	4.52
	Too few						
SF	6.06	4.53	7.59	3.06	4.52	7.59	3.07
SFRR	5.39	4.15	6.63	2.48	3.87	7.49	3.62
CAL	5.62	4.50	6.74	2.25	3.73	7.51	3.78
FWA	5.19	3.99	6.39	2.40	4.38	7.62	3.24
HAR	5.34	4.22	6.46	2.23	4.38	7.47	3.09
PML	5.33	4.09	6.56	2.47	3.74	7.35	3.62
PEL	5.49	4.27	6.72	2.45	4.51	6.63	2.12

Table A1.7: Point and 95% confidence level estimation of proportions using several methods for variance estimation. Main variable: "Confidence in immigrants"

In general, how much confidence do you have in immigrants?							
		Jackknife variance			Analytical variance		
Estimator	PROP	LB	UB	LEN	LB	UB	LEN
	None at all						
SF	7.15	5.56	8.75	3.18	5.56	8.75	3.19
SFRR	7.66	6.07	9.24	3.18	5.59	9.36	3.77
CAL	7.17	5.91	8.43	2.51	5.20	9.14	3.93
FWA	7.15	5.67	8.64	2.98	5.47	8.88	3.41
HAR	7.01	5.70	8.31	2.61	5.48	8.68	3.20
PML	7.36	5.76	8.97	3.21	5.87	9.80	3.93
PEL	7.14	5.73	8.54	2.81	7.14	9.37	2.23
	Very little						
SF	35.67	32.43	38.90	6.47	32.46	38.88	6.42
SFRR	34.61	31.83	37.40	5.56	31.46	38.42	6.96
CAL	34.34	32.04	36.65	4.61	30.71	37.98	7.27
FWA	34.09	31.16	37.02	5.86	32.80	39.46	6.66
HAR	33.65	31.01	36.29	5.28	32.37	38.80	6.43
PML	34.44	31.36	37.52	6.16	32.32	39.85	7.53
PEL	34.71	32.09	37.32	5.22	34.14	38.33	4.19
	Quite a lot						
SF	35.02	32.06	37.98	5.92	32.09	37.94	5.85
SFRR	36.45	33.80	39.10	5.31	32.36	38.98	6.62
CAL	36.55	34.27	38.84	4.57	33.10	40.01	6.91
FWA	38.18	35.14	41.23	6.09	32.08	38.21	6.13
HAR	38.12	35.39	40.84	5.45	32.07	37.96	5.89
PML	37.34	34.63	40.05	5.42	31.80	38.72	6.92
PEL	37.06	34.44	39.67	5.22	32.72	37.06	4.34
	Very much						
SF	12.24	10.13	14.34	4.20	10.13	14.34	4.21
SFRR	10.94	9.30	12.59	3.28	8.85	13.75	4.90
CAL	11.34	9.82	12.86	3.05	8.78	13.90	5.12
FWA	10.80	9.10	12.50	3.40	9.95	14.40	4.45
HAR	10.90	9.36	12.44	3.08	9.94	14.16	4.22
PML	10.90	9.22	12.57	3.35	8.58	13.56	4.98
PEL	11.18	9.48	12.88	3.40	9.36	12.10	2.74
	It depends						
SF	7.29	5.73	8.85	3.13	5.73	8.85	3.12
SFRR	6.87	5.71	8.04	2.33	5.64	9.33	3.69
CAL	7.63	6.38	8.87	2.49	5.70	9.56	3.85
FWA	6.68	5.42	7.94	2.51	5.18	8.43	3.25
HAR	7.25	6.03	8.47	2.44	5.17	8.27	3.10
PML	6.71	5.53	7.90	2.37	4.72	8.33	3.61
PEL	7.05	5.73	8.36	2.63	5.27	8.27	3.00

Table A1.8: Frame sizes

	ICT-H 2012	*ICT-H 2013*	*IESA-SH 2012*
N_A	4982920	4507662	4880574
N_B	5707655	6073789	6098453
N_{ab}	4339659	3983443	4266797

IESA-SH, Survey in Households, IESA
ICT-H, Survey on the Equipment and Use of Information and Communication Technologies in Households, INE

Table A1.9: Average, AVG and coefficient of variation, CV, of four point estimations. Main variable: "Goodness of immigration"

In general, do you think that for Andalusia, immigration is ...?					
Estimator	AVG	*ICT-H 2012*	*ICT-H 2013*	*IESA-SH 2012*	CV
			Very bad		
SFRR	13.96	13.90	13.95	14.04	0.51
CAL	13.45	13.35	13.47	13.54	0.71
PML	13.85	13.87	13.82	13.85	0.18
PEL	13.71	13.62	13.72	13.78	0.59
			Bad		
SFRR	47.24	47.39	47.19	47.14	0.28
CAL	47.02	46.92	47.05	47.08	0.18
PML	46.28	46.43	46.10	46.32	0.36
PEL	46.10	45.95	46.14	46.22	0.30
			Neither good nor bad		
SFRR	4.51	4.47	4.52	4.53	0.71
CAL	4.77	4.75	4.77	4.80	0.53
PML	4.40	4.34	4.46	4.41	1.37
PEL	4.38	4.33	4.38	4.43	1.14
			Good		
SFRR	28.26	28.16	28.31	28.30	0.30
CAL	28.80	28.98	28.76	28.66	0.57
PML	28.06	27.77	27.86	28.55	1.52
PEL	29.72	29.96	29.67	29.53	0.74
			Very good		
SFRR	2.12	2.10	2.13	2.14	0.98
CAL	2.17	2.16	2.17	2.17	0.27
PML	2.11	2.08	2.13	2.11	1.19
PEL	2.12	2.12	2.12	2.13	0.27

Table A1.10: Average, AVG and coefficient of variation, CV, of four point estimations. Main variables: "amount immigration", "confidence in immigrants"

In relation to the number of immigrants currently living in Andalusia, do you think there are ...?					
Estimator	AVG	*ICT-H 2012*	*ICT-H 2013*	*IESA-SH 2012*	CV
		Too many			
SFRR	40.82	40.69	40.81	40.96	0.33
CAL	41.34	40.97	41.38	41.68	0.86
PML	40.46	40.44	40.46	40.53	0.12
PEL	41.42	41.05	41.45	41.75	0.85
		A reasonable number			
SFRR	47.86	47.85	47.88	47.86	0.03
CAL	46.69	47.03	46.65	46.39	0.69
PML	48.03	47.95	48.10	48.03	0.16
PEL	46.40	46.72	46.36	46.11	0.66
		Too few			
SFRR	5.44	5.39	5.44	5.49	0.92
CAL	5.74	5.62	5.75	5.85	2.01
PML	5.38	5.33	5.39	5.41	0.77
PEL	5.62	5.49	5.63	5.74	2.23
In general, how much confidence do you have in immigrants?					
Estimator	AVG	*ICT-H 2012*	*ICT-H 2013*	*IESA-SH 2012*	CV
		None at all			
SFRR	7.58	7.66	7.56	7.53	0.90
CAL	7.17	7.17	7.18	7.16	0.14
PML	7.27	7.36	7.16	7.28	1.39
PEL	7.14	7.14	7.15	7.13	0.14
		Very little			
SFRR	34.73	34.61	34.70	34.87	0.38
CAL	34.71	34.34	34.76	35.04	1.01
PML	34.46	34.44	34.34	34.61	0.40
PEL	35.06	34.71	35.10	35.36	0.93
		Quite a lot			
SFRR	36.45	36.45	36.49	36.40	0.12
CAL	36.12	36.55	36.06	35.75	1.12
PML	37.51	37.34	37.61	37.58	0.39
PEL	36.59	37.06	36.53	36.19	1.20
		Very much			
SFRR	11.14	10.94	11.16	11.32	1.71
CAL	11.59	11.34	11.60	11.82	2.07
PML	11.09	10.90	11.15	11.22	1.52
PEL	11.44	11.18	11.45	11.68	2.19
		It depends			
SFRR	6.73	6.87	6.73	6.58	2.16
CAL	7.53	7.63	7.52	7.45	1.20
PML	6.74	6.71	6.68	6.82	1.09
PEL	6.99	7.05	6.98	6.93	0.86

Appendix A2

Frames2: A package for estimation in dual frame surveys

Arcos, A., Molina, D., Ranalli, M. G. and Rueda, M. (2015)
Frames2: A package for estimation in dual frame surveys.
The R Journal, Vol. 7, Number 1, pp. 52 - 72

Abstract

Data from complex survey designs require special consideration with regard to estimation for finite population parameters and corresponding variance estimation procedures, as a consequence of significant departures from simple random sampling assumption. In the past decade a number of statistical software packages have been developed to facilitate the analysis of complex survey data. All these statistical software are able to treat samples selected from one sampling frame containing all population units. Dual frame surveys are very useful when it is not possible to guarantee a complete coverage of the target population and may result in considerable cost savings over a single frame design with comparable precision. There are several available estimators in the statistical literature but no existing software covers dual frame estimation procedures. This gap is now filled by *Frames2*. In this paper we highlight the main features of the package. The package includes the main estimators in dual frame surveys and

also provides interval confidence estimation.

A2.1 Introduction

Classic sampling theory usually assumes the existence of one sampling frame containing all finite population units. Then, a probability sample is drawn according to a given sampling design and information collected is used for estimation and inference purposes. In traditional design-based inference the population data are regarded as fixed and the randomness comes entirely from the sampling procedure. The most used design-based estimator is the Horvitz-Thompson estimator that is unbiased for the population total if the sampling frame includes all population units, if all sampled units respond and if there is no measurement error. In the presence of auxiliary information, there exist several procedures to obtain more efficient estimators for population means and totals of variable of interest; in particular, customary ratio, regression, raking, post-stratified and calibration estimators. Several software packages have been developed to facilitate the analysis of complex survey data and implement some of these estimators as SAS, SPSS, Systat, Stata, SUDAAN or PCCarp. CRAN contains several R packages that include these design-based methods typically used in survey methodology to treat samples selected from one sampling frame (e.g. *survey* (Lumley, 2014), *sampling* (Tillè and Matei, 2012), *laeken* (Alfons *et al.*, 2014) or *TeachingSampling* (Gutierrez Rojas, 2014) among others). Templ (2014) performs a detailed list of packages that includes methods to analyse complex surveys.

In practice, the assumption that the sampling frame contains all population units is rarely met. Often, one finds that sampling from a frame which is known to cover approximately all units in the population is quite expensive while other frames (e.g. special lists of units) are available for cheaper sampling methods. However, the latter usually only cover an unknown or only approximately known fraction of the population. A common example of frame undercoverage is provided by telephone surveys. Estimation could be affected by serious bias due to the lack of a telephone in some households and the generalised use of mobile phones, which are sometimes replacing fixed (land) lines entirely. The potential for coverage error as a result of the exponential growth of the cell-phone only population has led to the development of dual-frame surveys. In these designs, a traditional sample from the landline frame is supplemented with an independent sample from a register of cell-phone numbers.

Dual frame sampling approach assumes that two frames are available for sampling and that, overall,

they cover the entire target population. The most common situation is the one represented in Figure A2.1 where the two frames, say frame A and frame B, show a certain degree of overlapping, so it is possible to distinguish three disjoint non-empty domains: domain a, containing units belonging to frame A but not to frame B; domain b, containing units belonging to frame B but not to frame A and domain ab, containing units belonging to both frames. As an example, consider a telephone survey where both landline and cell phone lists are available; let A be the landline frame and B the cell phone frame. Then, it is possible to distinguish three types of individuals: landline only units, cell-only units and units with both landline and cell phone, which will compose domain a, b and ab, respectively.

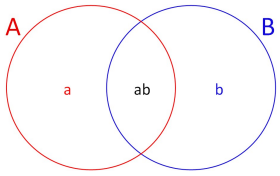


Figure A2.1: Two frames with overlapping.

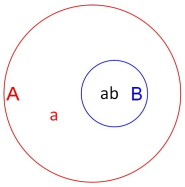


Figure A2.2: Frame B is included in frame A.

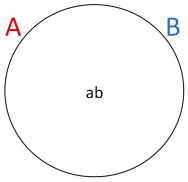


Figure A2.3: Frame A and frame B exactly match.

Nevertheless, one can face some other situations depending on the relative positions of the frames. For example, Figure A2.2 shows the case in which frame B is totally included in frame A, that is, frame B is a subset of frame A. Here domain b is empty. We also may find scenarios where the two sampling frames exactly match, as depicted in Figure A2.3, where ab is the only non-empty domain. Finally, the scenario where domain ab is empty has no interest from a dual frame perspective, since it can be considered as a special case of stratified sampling.

Whatever the scenario, an appropriate choice of the frames results in a better coverage of the target population, which, in turn, leads to a better efficiency of estimators calculated from data from dual frame surveys. This point is particularly important when estimating parameters in rare or elusive populations, where undercoverage errors are usually due to the difficulty of finding individuals showing the characteristic under study when sampling from only one general frame. This issue can be dealt with by incorporating a second frame with a high density of members of the rare population so that the two frames are, together, now complete. Dual frame sampling as a method of improvement of efficiency may seem expensive and unviable, but it is not. In fact, Hartley (1962) notes that dual frame surveys can result in important cost savings in comparison with single frame ones with a comparable efficiency. As

an additional interesting characteristic, dual frame methodology offers the researcher the possibility to consider different data collection procedures and/or different sampling designs, one for each frame. Dual frame surveys have gained much attention and became largely used by statistical agencies and private organizations to take advantage of these benefits.

Standard software packages for complex surveys can not be used directly when the sample is obtained from a dual frame survey because the classical design-based estimators are severely biased and there is a underestimation of standard errors. Weighted analyses with standard statistical software, with certain modified weights, can yield correct point estimates of population parameters but still yield incorrect results for estimated standard errors. A number of authors have developed methods for estimating population means and totals from dual (or, more generally, multiple) frame surveys but most of these methods require ad-hoc software for their implementation. To the best of our knowledge, there is no software incorporating these estimation procedures for handling dual frame surveys.

Frames2 (Arcos *et al.*, 2015) tries to fill this gap by providing functions for point and interval estimation from dual frame surveys. The paper is organized as follows. In the next section, we provide an overview of the main point estimators proposed so far in the dual frame context and reviews also jackknife variance estimation as a tool to compare efficiency for all of them. Subsequently, we present package *Frames2*, discussing guidelines that have been followed to construct it and presenting its principal functions and functionalities. We also provide examples to illustrate how the package works.

A2.2 Estimation in dual frame surveys

Consider again the situation depicted in Figure A2.1. Assume we have a finite set of N population units identified by integers, $\mathcal{U} = \{1, \ldots, k, \ldots, N\}$, and let A and B be two sampling-frames, both can be incomplete, but it is assumed that together they cover the entire finite population. Let $\mathcal{A}$ be the set of population units in frame A and $\mathcal{B}$ the set of population units in frame B. The population of interest, $\mathcal{U}$, may be divided into three mutually exclusive domains, $a = \mathcal{A} \cap \mathcal{B}^c, b = \mathcal{A}^c \cap \mathcal{B}$ and $ab = \mathcal{A} \cap \mathcal{B}$. Let N, N_A, N_B, N_a, N_b and N_{ab} be the number of population units in $\mathcal{U}, \mathcal{A}, \mathcal{B}, a, b, ab$, respectively.

Let y be a variable of interest in the population and let y_k be its value on unit k, for $k = 1, ..., N$.

The objective is to estimate the finite population total $Y = \sum_k y_k$ that can be written as

$$Y = Y_a + Y_{ab} + Y_b,$$

where $Y_a = \sum_{k\in a} y_k, Y_{ab} = \sum_{k\in ab} y_k$ and $Y_b = \sum_{k\in b} y_k$. To this end, independent samples s_A and s_B are drawn from frame A and frame B of sizes n_A and n_B, respectively. Unit k in $\mathcal{A}$ has first-order inclusion probability $\pi_k^A = Pr(k \in s_A)$ and unit k in $\mathcal{B}$ has first-order inclusion probability $\pi_k^B = Pr(k \in s_B)$.

From data collected in s_A, it is possible to compute one unbiased estimator of the total for each domain in frame A, $\hat{Y}_a$ and $\hat{Y}_{ab}^A$, as described below:

$$\hat{Y}_a = \sum_{k\in s_A} \delta_k(a) d_k^A y_k, \qquad \hat{Y}_{ab}^A = \sum_{k\in s_A} \delta_k(ab) d_k^A y_k,$$

where $\delta_k(a) = 1$ if $k \in a$ and 0 otherwise, $\delta_k(ab) = 1$ if $k \in ab$ and 0 otherwise and d_k^A are the weights under the sampling design used in frame A, defined as the inverse of the first order inclusion probabilities, $d_k^A = 1/\pi_k^A$. Similarly, using information included in s_B, one can obtain an unbiased estimator of total for domain b and another one for domain ab, $\hat{Y}_b$ and $\hat{Y}_{ab}^B$, which can be expressed as

$$\hat{Y}_b = \sum_{k\in s_B} \delta_k(b) d_k^B y_k, \qquad \hat{Y}_{ab}^B = \sum_{k\in s_B} \delta_k(ab) d_k^B y_k,$$

with $\delta_k(b) = 1$ if $k \in b$ and 0 otherwise, and d_k^B the weights under the sampling design used in frame B defined as the inverse of the first order inclusion probabilities, $d_k^B = 1/\pi_k^B$.

Different approaches for estimating the population total from dual frame surveys have been proposed in the literature. Hartley (1962) suggests the use of a parameter, θ, to weight $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$, providing the estimator

$$\hat{Y}_H = \hat{Y}_a + \theta\hat{Y}_{ab}^A + (1-\theta)\hat{Y}_{ab}^B + \hat{Y}_b, \tag{A2.1}$$

where $\theta \in [0, 1]$. Hartley (1974) himself proved that

$$\theta_{opt} = \frac{V(\hat{Y}_{ab}^B) + Cov(\hat{Y}_b, \hat{Y}_{ab}^B) - Cov(\hat{Y}_a, \hat{Y}_{ab}^A)}{V(\hat{Y}_{ab}^A) + V(\hat{Y}_{ab}^B)}$$

is the optimum value for θ so that variance of the estimator with respect to the design is minimized. In

practice, θ_{opt} cannot be computed, since population variances and covariances involved in its calculation are unknown, so they must be estimated from sampling data. An estimator for the variance of $\hat{Y}_H$ can be computed, taking into account that samples from frame A and frame B are drawn independently, as follows

$$\hat{V}(\hat{Y}_H) = \hat{V}(\hat{Y}_a) + \theta^2\hat{V}(\hat{Y}^A_{ab}) + \theta\widehat{Cov}(\hat{Y}_a, \hat{Y}^A_{ab}) + (1-\theta)^2\hat{V}(\hat{Y}^B_{ab}) + \hat{V}(\hat{Y}_b) + (1-\theta)\widehat{Cov}(\hat{Y}_b, \hat{Y}^B_{ab}), \quad \text{(A2.2)}$$

where hats denote suitable variance and covariance estimators.

Fuller and Burmeister (1972) introduce information from the estimation of overlap domain size, obtaining the following estimator

$$\hat{Y}_{FB} = \hat{Y}_a + \hat{Y}_b + \beta_1\hat{Y}^A_{ab} + (1-\beta_1)\hat{Y}^B_{ab} + \beta_2(\hat{N}^A_{ab} - \hat{N}^B_{ab}), \quad \text{(A2.3)}$$

where $\hat{N}^A_{ab} = \sum_{k\in s_A}\delta_k(ab)d^A_k$ and $\hat{N}^B_{ab} = \sum_{k\in s_B}\delta_k(ab)d^B_k$. Fuller and Burmeister (1972) also show that

$$\begin{bmatrix}\tilde{\beta}_1\\ \tilde{\beta}_2\end{bmatrix} = -\begin{bmatrix} V(\hat{Y}^A_{ab} - \hat{Y}^B_{ab}) & Cov(\hat{Y}^A_{ab} - \hat{Y}^B_{ab}, \hat{N}^A_{ab} - \hat{N}^B_{ab})\\ Cov(\hat{Y}^A_{ab} - \hat{Y}^B_{ab}, \hat{N}^A_{ab} - \hat{N}^B_{ab}) & V(\hat{N}^A_{ab} - \hat{N}^B_{ab})\end{bmatrix}^{-1} \times \begin{bmatrix} Cov(\hat{Y}_a + \hat{Y}_b + \hat{Y}^B_{ab}, \hat{Y}^A_{ab} - \hat{Y}^B_{ab})\\ Cov(\hat{Y}_a + \hat{Y}_b + \hat{Y}^B_{ab}, \hat{N}^A_{ab} - \hat{N}^B_{ab})\end{bmatrix}$$

are the optimal values for β_1 and β_2 in the sense that they minimize the variance of the estimator. Again, $\tilde{\beta}_1$ and $\tilde{\beta}_2$ need to be estimated, since population values are not known in practice. An estimator for the variance of $\hat{Y}_{FB}$ is given by

$$\hat{V}(\hat{Y}_{FB}) = \hat{V}(\hat{Y}_a) + \hat{V}(\hat{Y}_B) + \hat{\beta}_1(\widehat{Cov}(\hat{Y}_a, \hat{Y}^A_{ab}) - \widehat{Cov}(\hat{Y}_B, \hat{Y}^B_{ab})) + \hat{\beta}_2(\widehat{Cov}(\hat{Y}_a, \hat{N}^A_{ab}) - \widehat{Cov}(\hat{Y}_B, \hat{N}^B_{ab})), \quad \text{(A2.4)}$$

with $\hat{Y}_B = \hat{Y}_b + \hat{Y}^B_{ab}$.

Bankier (1986) and Kalton and Anderson (1986) combine all sampling units coming from the two frames, s_A and s_B, trying to build a single sample as if it was drawn from only one frame. Sampling weights for the units in the overlap domain need, then, to be modified to avoid bias. These adjusted

weights are

$$\tilde{d}_k^A = \begin{cases} d_k^A & \text{if } k \in a \\ (1/d_k^A + 1/d_k^B)^{-1} & \text{if } k \in ab \end{cases} \qquad \text{and} \qquad \tilde{d}_k^B = \begin{cases} d_k^B & \text{if } k \in b \\ (1/d_k^A + 1/d_k^B)^{-1} & \text{if } k \in ab \end{cases}$$

or, summarizing,

$$\tilde{d}_k = \begin{cases} d_k^A & \text{if } k \in a \\ (1/d_k^A + 1/d_k^B)^{-1} & \text{if } k \in ab \\ d_k^B & \text{if } k \in b \end{cases}. \tag{A2.5}$$

Hence, the estimator can be expressed in the form

$$\hat{Y}_{BKA} = \sum_{k \in s_A} \tilde{d}_k^A y_k + \sum_{k \in s_B} \tilde{d}_k^B y_k = \sum_{k \in s} \tilde{d}_k y_k, \tag{A2.6}$$

with $s = s_A \cup s_B$. Note that to compute this estimator, one needs to know, for units in sample coming from the overlap domain, the inclusion probability under both sampling designs. Rao and Skinner (1996) propose the following unbiased estimator for the variance of the estimator

$$\hat{V}(\hat{Y}_{BKA}) = \hat{V}(\sum_{k \in s_A} \tilde{z}_k^A) + \hat{V}(\sum_{k \in s_B} \tilde{z}_k^B), \tag{A2.7}$$

where $\tilde{z}_k^A = \delta_k(a) y_k + (1 - \delta_k(a)) y_k \frac{\pi_k^A}{\pi_k^A + \pi_k^B}$ and $\tilde{z}_k^B = \delta_k(b) y_k + (1 - \delta_k(b)) y_k \frac{\pi_k^B}{\pi_k^A + \pi_k^B}$.

When frame sizes, N_A and N_B, are known, estimator (A2.6) can be adjusted to increase efficiency through different procedures as, for example, raking ratio (Bankier, 1986; Skinner, 1991). Applying the latter, one obtains a new estimator, usually called raking ratio (Skinner, 1991), which has the form

$$\hat{Y}_{SFRR} = \frac{N_A - \hat{N}_{ab}^{rake}}{\hat{N}_a} \hat{Y}_a^A + \frac{N_B - \hat{N}_{ab}^{rake}}{\hat{N}_b} \hat{Y}_b^B + \frac{\hat{N}_{ab}^{rake}}{\hat{N}_{abS}} \hat{Y}_{abS}, \tag{A2.8}$$

where $\hat{Y}_{abS} = \sum_{k \in s_A} \tilde{d}_k^A \delta_k(ab) y_k + \sum_{k \in s_B} \tilde{d}_k^B \delta_k(ab) y_k$, $\hat{N}_{abS} = \sum_{k \in s_A} \tilde{d}_k^A \delta_k(ab) + \sum_{k \in s_B} \tilde{d}_k^B \delta_k(ab)$, $\hat{N}_a = \sum_{k \in s_A} \delta_k(a)$, $\hat{N}_b = \sum_{k \in s_B} \delta_k(b)$ and $\hat{N}_{ab}^{rake}$ is the smaller root of quadratic equation $\hat{N}_{abS} x^2 - (\hat{N}_{abS}(N_A + N_B) + \hat{N}_{aS}^A \hat{N}_{bS}^B) x + \hat{N}_{abS} N_A N_B = 0$.

Skinner and Rao (1986) use a pseudo maximum likelihood approach to extend to complex designs the maximum likelihood estimator proposed by Fuller and Burmeister (1972) only for simple random

sampling without replacement. The resulting estimator is given by

$$\begin{aligned}\hat{Y}_{PML} = &\frac{N_A - \hat{N}_{ab}^{PML}(\gamma)}{\hat{N}_a^A}\hat{Y}_a^A + \frac{N_B - \hat{N}_{ab}^{PML}(\gamma)}{\hat{N}_b^B}\hat{Y}_b^B \\ &+ \frac{\hat{N}_{ab}^{PML}(\gamma)}{\gamma\hat{N}_{ab}^A + (1-\gamma)\hat{N}_{ab}^B}[\gamma\hat{Y}_{ab}^A + (1-\gamma)\hat{Y}_{ab}^B],\end{aligned} \tag{A2.9}$$

where $\hat{N}_{ab}^{PML}(\gamma)$ is the smallest of the roots of quadratic equation $[\gamma/N_B+(1-\gamma)/N_A]x^2-[1+\gamma\hat{N}_{ab}^A/N_B+(1-\gamma)\hat{N}_{ab}^B/N_A]x+\gamma\hat{N}_{ab}^A+(1-\gamma)\hat{N}_{ab}^B=0$ and $\gamma\in(0,1)$. It is also shown that the following value for γ

$$\gamma_{opt} = \frac{\hat{N}_a N_B V(\hat{N}_{ab}^B)}{\hat{N}_a N_B V(\hat{N}_{ab}^B) + \hat{N}_b N_A V(\hat{N}_{ab}^A)} \tag{A2.10}$$

minimizes the variance of $\hat{Y}_{PML}$. One can use the delta method to obtain a consistent estimator of the variance of this estimator in the form

$$\hat{V}(\hat{Y}_{PML}) = \hat{V}(\sum_{k\in s_A}\tilde{z}_k^A) + \hat{V}(\sum_{k\in s_B}\tilde{z}_k^B), \tag{A2.11}$$

where, in this case, $\tilde{z}_k^A = y_k - \frac{\hat{Y}_a}{\hat{N}_a}$ if $k\in a$ and $\tilde{z}_k^A = \hat{\gamma}_{opt}\left(y_k - \frac{\hat{Y}_{ab}^A}{\hat{N}_{ab}^A}\right) + \hat{\lambda}\hat{\phi}$ if $k\in ab$, with $\hat{\gamma}_{opt}$ an estimator of γ_{opt} in (A2.10) obtained by replacing population quantities with their estimators, $\hat{\lambda} = \frac{n_A/N_A\hat{Y}_{ab}^A + n_B/N_B\hat{Y}_{ab}^B}{n_A/N_A\hat{N}_{ab}^A + n_B/N_B\hat{N}_{ab}^B} - \frac{\hat{Y}_a}{\hat{N}_a} - \frac{\hat{Y}_b}{\hat{N}_b}$ and $\hat{\phi} = \frac{n_A\hat{N}_b}{n_A\hat{N}_b + n_B\hat{N}_a}$. Similary, one can define $\tilde{z}_k^B = y_k - \frac{\hat{Y}_b}{\hat{N}_b}$ if $k\in b$ and $\tilde{z}_k^B = (1-\hat{\gamma}_{opt})\left(y_k - \frac{\hat{Y}_{ab}^B}{\hat{N}_{ab}^B}\right) + \hat{\lambda}(1-\hat{\phi})$ if $k\in ab$.

More recently, Rao and Wu (2010) have proposed a pseudo empirical likelihood estimator for the population mean based on poststratified samples. Such estimator is computed as

$$\hat{\bar{Y}}_{PEL} = \frac{N_a}{N}\hat{\bar{Y}}_a + \frac{\eta N_{ab}}{N}\hat{\bar{Y}}_{ab}^A + \frac{(1-\eta)N_{ab}}{N}\hat{\bar{Y}}_{ab}^B + \frac{N_b}{N}\hat{\bar{Y}}_b, \tag{A2.12}$$

where, in this case, $\hat{\bar{Y}}_a = \sum_{k\in s_A}\hat{p}_{ak}y_k\delta_k(a)$, $\hat{\bar{Y}}_{ab}^A = \sum_{k\in s_A}\hat{p}_{abk}^A y_k\delta_k(ab)$, $\hat{\bar{Y}}_{ab}^B = \sum_{k\in s_B}\hat{p}_{abk}^B y_k\delta_k(ab)$ and $\hat{\bar{Y}}_b = \sum_{k\in s_B}\hat{p}_{bk}y_k\delta_k(b)$ with $\hat{p}_{ak}$, $\hat{p}_{abk}^A$, $\hat{p}_{abk}^B$ and $\hat{p}_{bk}$ the weights resulting from maximizing the pseudo empirical likelihood procedure under a set of constraints (see Rao and Wu (2010) for details). Furthermore, $\eta\in(0,1)$. In this case, it is assumed that N_A, N_B and N_{ab} are known, but this is not always the case. Authors also provide modifications to be carried out in (A2.12) to adapt it to situations where only N_A and N_B are known or where none of N_A, N_B or N_{ab} are known. In addition, auxiliary information

coming from either one or both frames can be incorporated to the estimation process to improve the accuracy of the estimates. In addition, instead of an analytic form for the variance of this estimator, Rao and Wu (2010) propose to compute confidence intervals using the bi-section method described by Wu (2005) for one single frame and extending it to the dual frame case. This method constructs intervals in the form $\{\theta | r_{ns}(\theta) < \chi^2_1(\alpha)\}$, where $\chi^2_1(\alpha)$ is the $1-\alpha$ quantile from a χ^2 distribution with one degree of freedom and $r_{ns}(\theta)$ represents the so called pseudo empirical log likelihood ratio statistic, which can be obtained as a difference of two pseudo empirical likelihood functions.

Recently, Ranalli *et al.* (2013) extended calibration procedures to estimation from dual frame sampling assuming that some kind of auxiliary information is available. For example, assuming there are p auxiliary variables, $\mathbf{x}_k(x_{1k}, ..., x_{pk})$ is the value taken by such auxiliary variables on unit k. Each auxiliary variable may be available only for units in frame A, only for units in frame B or for units in the whole population. In addition, it is assumed that the vector of population totals of the auxiliary variables, $\mathbf{t_x} = \sum_{k \in \mathcal{U}} \mathbf{x}_k$ is also known. In this context, the dual frame calibration estimator can be defined as follows

$$\hat{Y}_{CALDF} = \sum_{k \in s} d_k^{CALDF} y_k, \tag{A2.13}$$

where weights d_k^{CALDF} are such that $min \sum_{k \in s} G(d_k^{CALDF}, \breve{d}_k)$ subject to $\sum_{k \in s} d_k^{CALDF} \mathbf{x}_k = \mathbf{t_x}$, with $G(\cdot, \cdot)$ a determined distance measure and

$$\breve{d}_k = \begin{cases} d_k^A & \text{if } k \in a \\ \eta d_k^A & \text{if } k \in ab \bigcap s_A \\ (1-\eta) d_k^B & \text{if } k \in ab \bigcap s_B \\ d_k^B & \text{if } k \in b \end{cases}, \tag{A2.14}$$

being $\eta \in [0, 1]$.

Then, with a similar approach to that of $\hat{Y}_{BKA}$, another calibration estimator can be computed as

$$\hat{Y}_{CALSF} = \sum_{k \in s} d_k^{CALSF} y_k, \tag{A2.15}$$

with weights d_k^{CALSF} verifying that $min \sum_{k \in s} G(d_k^{CALSF}, \tilde{d}_k)$ subject to $\sum_{k \in s} d_k^{CALSF} \mathbf{x}_k = \mathbf{t_x}$, being $\tilde{d}_k$ the weights defined in (A2.5).

An estimator of the variance of any calibration estimator can be obtained using Deville's method

(Deville, 1993) through following expression

$$\hat{V}(\hat{Y}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s} \frac{d_k^\star - 1}{d_k^\star} \left(d_k^\star e_k - \sum_{l\in s} a_l d_l^\star e_l \right)^2, \tag{A2.16}$$

where $d_k^\star$ is given by (A2.5) or by (A2.14) according to whether we use $\hat{Y}_{CALSF}$ or $\hat{Y}_{CALDF}$, respectively. In addition, $a_k = \frac{d_k^\star - 1}{d_k^\star} / \sum_{l\in s} \frac{d_l^\star - 1}{d_l^\star}$ and e_k are the residuals of the generalized regression of y on $\mathbf{x}$.

Some of the estimators described above are particular types of calibration estimators. For example, estimator (A2.8) can be obtained as a particular case of $\hat{Y}_{CALSF}$ in the case where frame sizes N_A and N_B are known and the "raking" method is selected for calibration. Having noted this, one can use (A2.16) to calculate an estimator of variance of (A2.8). See Ranalli *et al.* (2013) for more details.

Table A2.1 shows a summary of the previous dual frame estimators according to the auxiliary information required. It can be noted that Hartley, FB and BKA estimators can be computed even when no information is available, but they cannot incorporate some auxiliary information when available. PML and SFRR can incorporate information on N_A and N_B, but PEL and CAL type estimators are the most flexible in that they can incorporate any kind of auxiliary information available.

Table A2.1: Estimator's capabilities versus auxiliary information availability

	None	N_A, N_B *known*	N_a, N_b *and* N_{ab} *known*	N_a, N_{ab}, N_b *and* X_A *and/or* X_B *known*
Hartley	✓			
FB	✓			
PML		✓		
PEL	✓	✓	✓	✓
CalDF	✓	✓	✓	✓
BKA*	✓			
SFRR*		✓		
CalSF*	✓	✓	✓	✓

(*) Inclusion probabilities are known in overlap domain ab for both frames

A2.2.1 Jackknife variance estimation

Variance estimation methods exposed so far depend on each specific estimator, so comparisons between variance estimations may lead to incorrect conclusions. Instead, one can consider jackknife, originally

proposed by Quenouille (1949, 1956) (see Wolter (2007) for a detailed description of this method in survey sampling) and extended to dual frame surveys by Lohr and Rao (2000), which can be used to estimate variances irrespective of the type of estimator allowing us to compare estimated efficiency for different estimators.

For a non stratified design in each frame, the jackknife estimator of the variance for any of the estimators described, generically denoted by $\hat{Y}_c$, is given by

$$v_J(\hat{Y}_c) = \frac{n_A - 1}{n_A} \sum_{i \in s_A} (\hat{Y}_c^A(i) - \overline{Y}_c^A)^2 + \frac{n_B - 1}{n_B} \sum_{j \in s_B} (\hat{Y}_c^B(j) - \overline{Y}_c^B)^2, \tag{A2.17}$$

with $\hat{Y}_c^A(i)$ the value of estimator $\hat{Y}_c$ after dropping unit i from s_A and $\overline{Y}_c^A$ the mean of values $\hat{Y}_c^A(i)$. Similarly, one can define $\hat{Y}_c^B(j)$ and $\overline{Y}_c^B$.

Jackknife may present an important bias when designs are without replacement. One could, then, incorporate an approximate finite-population correction to estimation to achieve unbiasedness. For example, assuming that a finite-population correction is needed in frame A, a modified jackknife estimator of variance, $v_J^*(\hat{Y}_c)$, can be calculated by replacing $\hat{Y}_c^A(i)$ in (A2.17) with $\hat{Y}_c^{A*}(i) = \hat{Y}_c + \sqrt{1 - \overline{\pi}_A}(\hat{Y}_c^A(i) - \hat{Y}_c)$, where $\overline{\pi}_A = \sum_{k \in s_A} \pi_k^A / n_A$.

Consider now a stratified design in each frame, where frame A is divided into H strata and frame B is divided into L strata. From stratum h of frame A, a sample of n_{Ah} units from the N_{Ah} population units in the stratum is drawn. Similarly, in stratum l of frame B, one selects n_{Bl} units from the N_{Bl} composing the stratum. Jackknife estimator of the variance can be defined, then, as follows

$$v_J(\hat{Y}_c) = \sum_{h=1}^{H} \frac{n_{Ah} - 1}{n_{Ah}} \sum_{i \in s_{Ah}} (\hat{Y}_c^A(hi) - \overline{Y}_c^{Ah})^2 + \sum_{l=1}^{L} \frac{n_{Bl} - 1}{n_{Bl}} \sum_{i \in s_{Bl}} (\hat{Y}_c^B(lj) - \overline{Y}_c^{Bl})^2, \tag{A2.18}$$

where $\hat{Y}_c^A(hi)$ is the value taken by $\hat{Y}_c$ after dropping unit i of stratum h from sample s_{Ah} and $\overline{Y}_c^{Ah}$ is the mean of values $\hat{Y}_c^A(hi)$. $\hat{Y}_c^B(lj)$ and $\overline{Y}_c^{Bl}$ can be defined in a similar way. Again, one can include an approximate finite-population correction in any stratum needing it. 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.

Stratified cluster sampling is very common in practice. Now we illustrate the jackknife estimator when a stratified sample of clusters is selected. Suppose that frame A has H strata and stratum h has

N_{Ah} observation units and $\widetilde{N}_{Ah}$ primary sampling units (clusters), of which $\widetilde{n}_{Ah}$ are sampled. Frame B has L strata, and stratum l has N_{Bl} observation units and $\widetilde{N}_{Bh}$ primary sampling units, of which $\widetilde{n}_{Bl}$ are sampled.

To define the jackknife estimator of the variance, let $\widetilde{Y}_c^A(hj)$ be the estimator of the same form as $\hat{Y}_c$ when the observations of sample primary sampling unit j of stratum h from sample in frame A are omitted. Similarly, $\widetilde{Y}_c^B(lk)$ is of the same form as $\hat{Y}_c$ when the observations of sample primary sampling unit k of stratum l from sample in frame B are omitted. The jackknife variance estimator is then given by:

$$v_J(\hat{Y}_c) = \sum_{h=1}^{H} \frac{\widetilde{n}Ah - 1}{\widetilde{n}_{Ah}} \sum_{j=1}^{\widetilde{n}_{Ah}} (\widetilde{Y}_c^A(hj) - \widetilde{Y}_c^{Ah})^2 + \sum_{l=1}^{L} \frac{\widetilde{n}_{Bl} - 1}{\widetilde{n}_{Bl}} \sum_{k \in s_{Bl}} (\widetilde{Y}_c^B(lk) - \widetilde{Y}_c^{Bl})^2, \tag{A2.19}$$

where $\widetilde{Y}_c^{Ah}$ is the mean of values $\widetilde{Y}_c^A(hj)$ and $\widetilde{Y}_c^{Bl}$ is the mean of values $\widetilde{Y}_c^B(lk)$.

A2.3 The R package *Frames2*

Frames2 is a new R package for point and interval estimation from dual frame sampling. It consists of eight main functions (`Hartley`, `FB`, `BKA`, `SFRR`, `PML`, `PEL`, `CalSF` and `CalDF`), each of them implementing one of the estimators described in the previous sections. The package also includes an additional function called `Compare` which provides a summary with all possible estimators that can be computed from the information provided as input. Moreover, six extra functions implementing auxiliary operations, like computation of Horvitz-Thompson estimators or of the covariance between two Horvitz-Thompson estimators, have also been included in the package to achieve a more understandable code. Finally, the package includes eight more functions, one for each estimator, for the calculation of confidence intervals based on the jackknife variance estimator.

A remarkable characteristic of these functions is the strong argument check. Functions check general aspects as the presence of `NA` or `NaN` values in its arguments, the number of main variables considered in the frames (that should match), the length of the arguments in each frame (that should also match) or the values for arguments indicating the domain each unit belongs to (which only can be "a" or "ab" for frame A or "b" or "ba" for frame B). If any issue is encountered, the function displays an error message

indicating what is the problem and what is the argument causing it, so that the user can manage errors easily. Furthermore, each function has additional checks depending on its specific characteristics or arguments. The main aim of this exhaustive check is to guarantee validity of the arguments, so one can avoid, to the extent possible, issues during computation.

Much attention has also been devoted to computational efficiency. Frequently, populations in a survey are extremely large or it is needed to keep sampling error below a certain value. As a consequence, one needs to consider large sample sizes, often in the order of tens of thousands sampling units. In these situations, computational efficiency of functions is essential, particularly when several variables are considered. Otherwise, user can face high runtimes and heavy computational loads. In this sense, functions of *Frames2* are developed according to strict efficiency measures, using the power of R to the matrix calculation to avoid loops and increase the computational efficiency. Table A2.2 shows user and system times necessary to compute estimators using an Intel(R) Core(TM) i7-3770 at 3.40 GHz when different sample sizes are considered. Elapsed time is also included to get an idea about the real time user needs to get estimations.

Table A2.2: User, system and elapsed times (in seconds) for estimators considering different sample sizes.

	user	system	elapsed
	$n_A = 10605, n_B = 13635$		
Hartley	0.01	0.02	0.04
FB	0.05	<0.01	0.07
BKA	0.03	<0.01	0.05
PML	0.02	0.02	0.03
SFRR	0.03	0.03	0.07
CalSF	0.03	<0.01	0.06
CalDF	0.04	0.01	0.05
	$n_A = 105105, n_B = 135135$		
Hartley	0.11	0.06	0.19
FB	0.27	0.07	0.32
BKA	0.13	0.05	0.17
PML	0.16	0.02	0.18
SFRR	0.42	0.12	0.54
CalSF	0.20	0.08	0.30
CalDF	0.22	0.07	0.31

Functions of *Frames2* have been implemented from an user-oriented perspective to increase usability. In this sense, most input parameters (which are the communication channel between the user and the function) are divided into two groups, depending on the frame they come from. This is to adapt functions as much as possible to the usual estimation procedure, in which the first step is to draw two independent samples, one from each frame. On the other hand, estimation details are managed internally by functions so that they are not visible for the user, who does not need to manage them.

Construction of functions has been carried out so that they perform properly in as many situations as possible. As noted in introductory section, one can face several situations when using two sampling frames depending on their relative positions. Although the most common is the one depicted in Figure A2.1, cases shown in Figures A2.2 and A2.3 may arise as well. All estimators described but PEL can be modified to cover these three situations, so corresponding functions of *Frames2* include necessary changes to produce estimates irrespective of the situation.

On the other hand, it is usual, when conducting a survey, to collect information on many variables of interest. To adapt to such situations, all functions are programmed to produce estimates when there are more than one variable of interest with only one call. To this end, parameters containing information about main variables observed in each frame can be either vectors, when only one variable is considered or matrices or data frames, when there are several variables under study. Cases in which the main aim of the survey is the estimation of population means or proportions are also very frequent. Hence, from the estimation of the population total for a variable, functions compute estimation of the mean as $\hat{\bar{Y}} = \hat{Y}/\hat{N}$. To obtain the estimation of the population size, functions internally apply the estimation procedure at issue to indicator vectors $\mathbf{1}_A$ and $\mathbf{1}_B$ of sizes n_A and n_B, respectively.

To get maximum flexibility, functions have been programmed to calculate estimates in cases in which user disposes of first and second order inclusion probabilities and in those other in which only first order ones are available, indistinctly. Knowledge of both first and second order inclusion probabilities is a strong assumption that does not always occur in practice. However, when calculating most of the estimators described in previous sections, second order inclusion probabilities are needed in many steps of the estimation procedure, mainly in computing estimated variances of a Horvitz-Thompson estimator or estimated covariances between two Horvitz-Thompson estimators. As an alternative, one can obtain variance estimations from only first order inclusion probabilities applying Deville's method reported in (A2.16), by substituting residuals e_k with the values of the variable of interest, y_k. Covariance estimations

are also obtained from variances through following expression

$$\widehat{Cov}(\hat{Y},\hat{X}) = \frac{\hat{V}(\hat{Y}+\hat{X}) - \hat{V}(\hat{Y}) - \hat{V}(\hat{X})}{2}.$$

To cover both cases, user has the possibility to consider different data structures for parameters relating to inclusion probabilities. So, if both first and second order inclusion probabilities are available, these parameters will be square matrices, whereas if only first order inclusion probabilities are known, these arguments will be vectors. The only restriction here is that type of both should match.

As can be deduced from previous sections, an essential aspect when computing estimates in dual frame is to know the domain each unit belongs to. Character vectors `domains_A` and `domains_B` are used for this purpose. The former can take values "a" or "ab", while the latter can take values "b" or "ba". Any other value will be considered as incorrect.

A2.3.1 Data description

To illustrate how functions operate, we use data sets `DatA` and `DatB`, both included in the package. `DatA` contains information about $n_A = 105$ households selected through a stratified sampling design from the $N_A = 1735$ households composing frame A. More specifically, frame A has been divided into 6 strata of sizes $N_{hA} = (727, 375, 113, 186, 115, 219)$ from which simple random without replacement samples of sizes $n_{hA} = (15, 20, 15, 20, 15, 20)$ have been drawn. On the other hand, a simple random without replacement sample of $n_B = 135$ households has been selected from the $N_B = 1191$ households in frame B. The size of the overlap domain for this case is $N_{ab} = 601$. This situation is depicted in Figure A2.4.

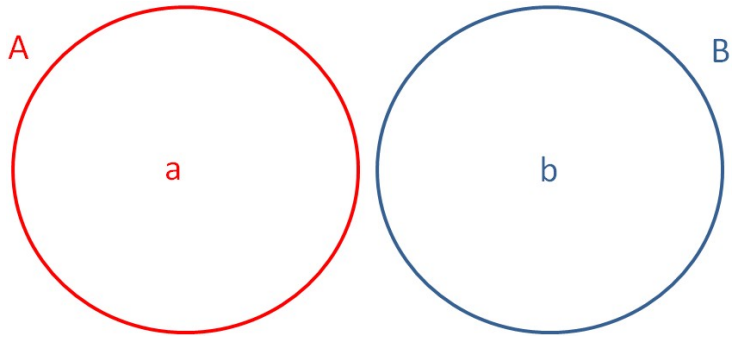


Figure A2.4: Frame and domain sizes for the data sets.

Both data sets contain information about the same variables. To better understand their structure, we report the first three rows of `DatA`:

```
> library (Frames2)
> data(DatA)
>
> head (DatA, 3)
  Domain   Feed   Clo   Lei     Inc    Tax     M2 Size      ProbA     ProbB Stratum
1      a 194.48 38.79 23.66 2452.07 112.90   0.00    0 0.02063274 0.0000000       1
2      a 250.23 16.92 22.68 2052.37 106.99   0.00    0 0.02063274 0.0000000       1
3     ab 199.95 24.50 23.24 2138.24 121.16 127.41    2 0.02063274 0.1133501       1
```

Each data set incorporates information about three main variables: Feeding, Clothing and Leisure. Additionally, there are two auxiliary variables for the units in frame A (Income and Taxes) and another two variables for units in frame B (Metres2 and Size). Corresponding totals for these auxiliary variables are assumed known in the entire frame and they are $T^A_{Inc} = 4300260, T^A_{Tax} = 215577, T^B_{M2} = 176553$ and $T^B_{Size} = 3529$. Finally, a variable indicating the domain each unit belongs to and two variables showing the first order inclusion probabilities for each frame complete the data sets.

Numerical square matrices `PiklA` and `PiklB`, with dimensions $n_A = 105$ and $n_B = 135$, are also used as probability inclusion matrices. These matrices contains second order inclusion probabilities and first order inclusion probabilities as diagonal elements. To check the appearance of these matrices let see the first submatrix of order 6 of `PiklA`.

```
> data(PiklA)
>
> PiklA[1:6, 1:6]
            [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
[1,] 0.020632737 0.000397876 0.000397876 0.000397876 0.000397876 0.000397876
[2,] 0.000397876 0.020632737 0.000397876 0.000397876 0.000397876 0.000397876
[3,] 0.000397876 0.000397876 0.020632737 0.000397876 0.000397876 0.000397876
[4,] 0.000397876 0.000397876 0.000397876 0.020632737 0.000397876 0.000397876
[5,] 0.000397876 0.000397876 0.000397876 0.000397876 0.020632737 0.000397876
```

```
[6,] 0.000397876 0.000397876 0.000397876 0.000397876 0.000397876 0.020632737
```

A2.3.2 No auxiliary information

When there is no further information than the one on the variables of interest, one can calculate some of the estimators described in previous section (as, for example, (A2.1) or (A2.3)) as follows

```
> library(Frames2)
>
> data(DatA)
> data(DatB)
> data(PiklA)
> data(PiklB)
>
> yA <- with(DatA, data.frame(Feed, Clo))
> yB <- with(DatB, data.frame(Feed, Clo))
>
> #Estimation for variables Feeding and Clothing using Hartley and Fuller-Burmeister
> #estimators with first and second order probabilities known
> Hartley(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain)

Estimation:
             Feed         Clo
Total 586959.9820 71967.62214
Mean     246.0429    30.16751
> FB(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain)

Estimation:
             Feed         Clo
Total 591665.5078 72064.99223
```

```
Mean     248.0153    30.20832
>
> #This is how estimates change when only first order probabilities are considered
> Hartley(yA, yB, DatA$ProbA, DatB$ProbB, DatA$Domain, DatB$Domain)

Estimation:
             Feed         Clo
Total 570867.8042 69473.86532
Mean     247.9484    30.17499
> FB(yA, yB, DatA$ProbA, DatB$ProbB, DatA$Domain, DatB$Domain)

Estimation:
             Feed         Clo
Total 571971.9511 69500.11448
Mean     248.4279    30.18639
```

As result, an object of class "`EstimatorDF`" is returned showing, by default, estimations for the population total and mean for the 2 considered variables. In general, m columns will be displayed, one for each of the m variables estimated. Further information about estimation process (as variance estimations or values of parameters involved in estimation) can be displayed by using function `summary`

```
> summary(Hartley(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain))

Call:
Hartley(ysA = yA, ysB = yB, pi_A = PiklA, pi_B = PiklB, domains_A = DatA$Domain,
    domains_B = DatB$Domain)

Estimation:
             Feed         Clo
Total 586959.9820 71967.62214
Mean     246.0429    30.16751
```

```
Variance Estimation:
                   Feed          Clo
Var. Total 2.437952e+08 4.728875e+06
Var. Mean  4.283804e+01 8.309261e-01

Total Domain Estimations:
                  Feed      Clo
Total dom. a  263233.1 31476.84
Total dom. ab 166651.7 21494.96
Total dom. b  164559.2 20451.85
Total dom. ba 128704.7 15547.49

Mean Domain Estimations:
                 Feed      Clo
Mean dom. a  251.8133 30.11129
Mean dom. ab 241.6468 31.16792
Mean dom. b  242.2443 30.10675
Mean dom. ba 251.5291 30.38466

Parameters:
           Feed       Clo
theta 0.8027766 0.7551851
```

Previous output shows in the component `Estimation` the estimations of the population total and the population mean computed using the Harley estimator, that is, $\hat{Y}_H$ and $\hat{\bar{Y}}_H$. Estimated variances of these estimations, $\hat{V}(\hat{Y}_H)$ and $\hat{V}(\hat{\bar{Y}}_H)$, are displayed in component `Variance Estimation`. In the section `Total Domain Estimations` we can see estimations $\hat{Y}_a, \hat{Y}^A_{ab}, \hat{Y}_b$ and $\hat{Y}^B_{ab}$. Estimates for the population mean for each domain, $\hat{\bar{Y}}_a, \hat{\bar{Y}}^A_{ab}, \hat{\bar{Y}}_b$ and $\hat{\bar{Y}}^B_{ab}$ are displayed in the component `Mean Domain Estimations`. Finally, $\hat{\theta}$, the estimated value of parameter involved in computation of the Hartley estimator is shown.

This additional information depends on the way each estimator is formulated. Thus, for example, extra information will include a parameter component when applied to a call to the Fuller-Burmeister estimator (and values of estimates for β_1 and β_2 will be displayed there), but not when applied to a call to the Bankier-Kalton-Anderson estimator (because no parameters are used when computing this estimator).

Results slightly change when a confidence interval is required. In that case, user has to indicate the confidence level desired for the interval through argument `conf_level` (default is `NULL`) and add it to the list of input parameters. The function calculates, then, a confidence interval based on the pivotal method. This method yields a confidence interval as follows: $\hat{Y} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{Y})}$ where $z_{\alpha/2}$ is the critical value of a standard normal distribution. Only for the case of PEL, confidence intervals are based on a χ^2 distribution and the bi-section method (Rao and Wu, 2010). In this case, default output will show 6 rows for each variable, lower and upper boundaries for confidence intervals are displayed together with estimates. So, one can obtain a 95% confidence interval for estimations in the last two of the previous four cases in this way

```
> Hartley(yA, yB, DatA$ProbA, DatB$ProbB, DatA$Domain, DatB$Domain, 0.95)

Estimation and  95 % Confidence Intervals:
                    Feed         Clo
Total       570867.8042 69473.86532
Lower Bound 511904.6588 61756.37677
Upper Bound 629830.9496 77191.35387
Mean           247.9484    30.17499
Lower Bound    222.3386    26.82301
Upper Bound    273.5582    33.52697
> FB(yA, yB, DatA$ProbA, DatB$ProbB, DatA$Domain, DatB$Domain, 0.95)

Estimation and  95 % Confidence Intervals:
                    Feed         Clo
Total       571971.9511 69500.11448
```

```
Lower Bound 513045.7170 61802.57411
Upper Bound 630898.1852 77197.65484
Mean           248.4279    30.18639
Lower Bound    222.8342    26.84307
Upper Bound    274.0217    33.52971
```

For estimators constructed as (A2.6), numeric vectors `pik_ab_B` and `pik_ba_A` of lengths n_A and n_B should be added as arguments. While `pik_ab_B` represents first order inclusion probabilities according to sampling design in frame B for units belonging to overlap domain selected in sample drawn from frame A, `pik_ba_A` contains first order inclusion probabilities according to sampling design in frame A for units belonging to overlap domain selected in sample drawn from frame B.

```
> yA <- with(DatA, data.frame(Feed, Clo, Lei))
> yB <- with(DatB, data.frame(Feed, Clo, Lei))
>
> #Bankier-Kalton-Anderson estimation and a 95% confidence
> #interval for the three main variables
> BKA(yA, yB, DatA$ProbA, DatB$ProbB, DatA$ProbB, DatB$ProbA, DatA$Domain,
+ DatB$Domain, 0.95)

Estimation and  95 % Confidence Intervals:
                   Feed         Clo         Lei
Total       566434.3200 68959.26705 50953.07583
Lower Bound 624569.2139 76538.11015 56036.23578
Upper Bound 508299.4262 61380.42395 45869.91588
Mean           247.8845    30.17814    22.29822
Lower Bound    273.3257    33.49482    24.52273
Upper Bound    222.4434    26.86147    20.07372
```

Note that these examples include just a few of the estimators that can be used when no auxiliary information is known. As noted in Table A2.1, other estimators, as those in (A2.12) or in (A3.8) or

in (A3.10), can be also calculated in this case.In this context, function `Compare` is quite useful, since it returns all possible estimators that can be computed according to the information provided as input.

```
> Compare(yA, yB, DatA$ProbA, DatB$ProbB, DatA$Domain, DatB$Domain)
$Hartley

Estimation:
             Feed         Clo        Lei
Total 570867.8042 69473.86532 51284.2727
Mean     247.9484    30.17499    22.2746

$FullerBurmeister

Estimation:
             Feed         Clo         Lei
Total 571971.9511 69500.11448 51210.03819
Mean     248.4279    30.18639    22.24236

$PEL

Estimation:
              Feed          Clo          Lei
Total 1.791588e+08 2.663164e+06 1.455533e+06
Mean  2.479314e+02 3.011373e+01 2.235969e+01

$Calibration_DF

Estimation:
             Feed         Clo        Lei
Total 595162.2604 72214.13351 53108.5059
```

```
Mean     248.8422    30.19332    22.2051
```

Using appropriate indicator variables as variables of interest, one can also estimate the overlap domain size, as shown below:

```
> indA <- as.integer(DatA$Domain == "ab")
> indB <- as.integer(DatB$Domain == "ba")
>
> Hartley(indA, indB, DatA$ProbA, DatB$ProbB, DatA$Domain, DatB$Domain)


Estimation:
             [,1]
Total 534.2743208
Mean    0.2320545
> BKA(indA, indB, DatA$ProbA, DatB$ProbB, DatA$ProbB, DatB$ProbA,
+ DatA$Domain, DatB$Domain)


Estimation:
             [,1]
Total 560.4121771
Mean    0.2452491
```

A2.3.3 Auxiliary information about frame sizes

For estimators requiring frame sizes known, as (A2.8) or (A2.9), it is needed to incorporate two additional input arguments, `N_A` and `N\_B`. There is also a group of estimators, including (A2.12) and (A3.10), that even being able to provide estimations without the need of auxiliary information, can use frame sizes to improve their precision. The following examples show the performance of these estimators.

```
> #SFRR estimator and CalSF estimator with frame sizes as auxiliary
> #information using method "raking" for the calibration for the three main variables
> SFRR (yA, yB, DatA$ProbA, DatB$ProbB, DatA$ProbB, DatB$ProbA, DatA$Domain,
```

```
+ DatB$Domain, N_A = 1735, N_B = 1191)

Estimation:
             Feed         Clo         Lei
Total 584713.4070 71086.18669 52423.74035
Mean     248.2219    30.17743    22.25487
> CalSF(yA, yB, DatA$ProbA, DatB$ProbB, DatA$ProbB, DatB$ProbA, DatA$Domain,
+ DatB$Domain, N_A = 1735, N_B = 1191, met = "raking")

Estimation:
             Feed         Clo         Lei
Total 584713.4070 71086.18669 52423.74035
Mean     248.2219    30.17743    22.25487
```

As highlighted previously, both results match. Note that argument `met` of SF calibration estimator indicates the method used in the calibration procedure. The possibility of choosing the calibration method is given by the fact that computation of both SF and DF calibration estimators is based on the function `calib` from package *sampling* (Till and Matei, 2012), which can manage three different calibration methods, each one associated with one particular distance measure. These methods are: linear, raking and logit.

Condition of knowing probabilities of inclusion in both frames for the units in the overlap domain may be restrictive is some cases. As an alternative, in cases where frame sizes are known but this condition is not met, it is possible to calculate dual frame estimators as (A2.9), (A2.12) or (A3.8). Next, it is illustrated how to obtain some of these estimators with *Frames2*.

```
> #Estimates for the three main variables using PML, PEL and CalDF
> #with frame sizes as auxiliary information in PEL and CalDF
> PML(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191)

Estimation:
             Feed         Clo         Lei
```

```
Total 593085.4467 72272.73759 53287.68044
Mean     248.0966    30.23277    22.29104
> PEL(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191)

Estimation:
             Feed         Clo         Lei
Total 590425.4843 72211.61334 53258.38286
Mean     247.4958    30.26982    22.32497
> CalDF(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191)

Estimation:
             Feed         Clo         Lei
Total 587502.4374 71368.45308 52490.98852
Mean     248.7193    30.21385    22.22207
```

To calculate `PEL` estimator, computational algorithms for the pseudo empirical likelihood method for the analysis of complex survey data presented by Wu (2005) have been used.

A2.3.4 Auxiliary information about domain sizes

In addition to the frame sizes, in some cases, it is possible to know the size of the overlap domain, N_{ab}. Generally, this highly improves the precision of the estimates. This situation has been taken into account when constructing functions implementing estimators (A2.12), (A3.8) and (A3.10), so user can incorporate this information through parameter `N_ab`, as shown below

```
> #Estimates for the three main variables using PEL estimator
> #with frame sizes and overlap domain size as auxiliary information
> PEL(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191,
+ N_ab = 601)

Estimation:
             Feed         Clo         Lei
```

```
Total 575289.2186 70429.95642 51894.32490
Mean     247.4362    30.29245    22.32014
> #Calibration estimators with the same auxiliary information
> #Estimates do not change when raking method is used for the calibration
> CalSF(yA, yB, PiklA, PiklB, DatA$ProbB, DatB$ProbA, DatA$Domain, DatB$Domain,
+ N_A = 1735, N_B = 1191, N_ab = 601)

Estimation:
             Feed         Clo         Lei
Total 577163.6066 70173.20412 51726.19862
Mean     248.2424    30.18202    22.24783
> CalSF(yA, yB, PiklA, PiklB, DatA$ProbB, DatB$ProbA, DatA$Domain, DatB$Domain,
+ N_A = 1735, N_B = 1191, N_ab = 601, met = "raking")

Estimation:
             Feed         Clo         Lei
Total 577163.6067 70173.20414 51726.19863
Mean     248.2424    30.18202    22.24783
> CalDF(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191,
+ N_ab = 601)

Estimation:
             Feed         Clo         Lei
Total 578691.1756 70246.32319 51600.78973
Mean     248.8994    30.21347    22.19389
> CalDF(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191,
+ N_ab = 601, met = "raking")

Estimation:
             Feed         Clo         Lei
```

```
Total 578691.1763 70246.32328 51600.78979
Mean     248.8994    30.21347    22.19389
```

Note that, in this case, calibration estimators provide the same results irrespective of the distance function employed. This is an interesting property that calibration estimators show only in the case in which all the domain sizes are known and used for calibration (see Deville, 1993).

A2.3.5 Auxiliary information about additional variables

On the other hand, some of the estimators are defined such that they can incorporate auxiliary information to the estimation process. This is the case of estimators (A2.12), (A3.8) and (A3.10). Functions implementing them are also able to manage auxiliary information. To achieve maximum flexibility, functions implementing estimators (A2.12), (A3.8) and (A3.10) are prepared to deal with auxiliary information when it is available only in frame A, only in frame B or in both frames. For instance, auxiliary information collected from frame A should be incorporated to functions through three arguments: `xsAFrameA` and `xsBFrameA`, numeric vectors, matrices or data frames (depending on the number of auxiliary variables in the frame); and `XA`, a numeric value or vector of length indicating population totals for the auxiliary variables considered in frame A. Similarly, auxiliary information in frame B is incorporated to each function through arguments `xsAFrameB`, `xsBFrameB` and `XB`. If auxiliary information is available in the whole population, it must be indicated through parameters `xsT` and `X`. In the following example, one can see how to calculate estimators using different type of auxiliary information

```
> #PEL, CalSF and CalDF estimators for the three main variables
> #using Income as auxiliary variable in frame A and Metres2 as auxiliary
> #variable in frame B assuming frame sizes known
> PEL(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191,
+ xsAFrameA = DatA$Inc, xsBFrameA = DatB$Inc, xsAFrameB = DatA$M2,
+ xsBFrameB = DatB$M2, XA = 4300260, XB = 176553)


Estimation:
             Feed         Clo         Lei
Total 587742.7193 71809.56826 53094.20112
```

```
Mean     246.3713    30.10129    22.25614
>
> CalSF(yA, yB, PiklA, PiklB, DatA$ProbB, DatB$ProbA, DatA$Domain, DatB$Domain,
+ N_A = 1735, N_B = 1191, xsAFrameA = DatA$Inc, xsBFrameA = DatB$Inc,
+ xsAFrameB = DatA$M2, xsBFrameB = DatB$M2, XA = 4300260, XB = 176553)

Estimation:
             Feed         Clo         Lei
Total 582398.3181 70897.88438 52252.24741
Mean     247.5819    30.13922    22.21282
>
> CalDF(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191,
+ xsAFrameA = DatA$Inc, xsBFrameA = DatB$Inc, xsAFrameB = DatA$M2,
+ xsBFrameB = DatB$M2, XA = 4300260, XB = 176553)

Estimation:
             Feed         Clo         Lei
Total 585185.4497 71194.61148 52346.43878
Mean     247.8075    30.14866    22.16705

> #Now, assume that overlap domain size is also known
> PEL(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191,
+ N_ab = 601, xsAFrameA = DatA$Inc, xsBFrameA = DatB$Inc,
+ xsAFrameB = DatA$M2, xsBFrameB = DatB$M2, XA = 4300260, XB = 176553)

Estimation:
             Feed         Clo         Lei
Total 572611.6997 69991.74803 51737.56089
Mean     246.2846    30.10398    22.25271
>
```

```
> CalSF(yA, yB, PiklA, PiklB, DatA$ProbB, DatB$ProbA, DatA$Domain, DatB$Domain,
+ N_A = 1735, N_B = 1191, N_ab = 601, xsAFrameA = DatA$Inc, xsBFrameA = DatB$Inc,
+ xsAFrameB = DatA$M2, xsBFrameB = DatB$M2, XA = 4300260, XB = 176553)

Estimation:
             Feed         Clo         Lei
Total 575636.7876 70076.78485 51628.27583
Mean     247.5857    30.14055    22.20571
>
> CalDF(yA, yB, PiklA, PiklB, DatA$Domain, DatB$Domain, N_A = 1735, N_B = 1191,
+ N_ab = 601, xsAFrameA = DatA$Inc, xsBFrameA = DatB$Inc,
+ xsAFrameB = DatA$M2, xsBFrameB = DatB$M2, XA = 4300260, XB = 176553)

Estimation:
             Feed        Clo         Lei
Total 576630.7609 70102.0037 51477.16737
Mean     248.0132    30.1514    22.14072
```

A2.3.6 Interval estimation based on jackknife variance estimation

Finally, eight additional functions have been included, each of them calculating confidence intervals based on jackknife variance estimator for each estimator. To carry out variance estimation using jackknife method, in addition to parameters to calculate each specific estimator, user has to indicate through arguments `sdA` and `sdB` the sampling design applied in each frame. Possible values are "srs" (simple random sampling without replacement), "str" (stratified sampling), "pps" (probabilities proportional to size sampling), "clu" (cluster sampling) or "strclu" (stratified cluster sampling). Default is "srs" for both frames. If a stratified or a cluster sampling has been carried out in one of the frames, it is needed to include information about the strata or the clusters. Furthermore, user is able to include a finite population correction factor in each frame by turning to `TRUE` parameters `fcpA` and `fcpB`, set by default to `FALSE`. Since main purpose of functions is to obtain confidence intervals, parameter `conf_level` is now

mandatory. As noted, these functions can be used, for example, to make comparisons between efficiency of estimators, as shown in next example.

```
> #Confidence intervals through jackknife for the three main variables
> #for estimators defined under the so called single frame approach with
> #a stratified random sampling in frame A and a simple random sampling
> #without replacement in frame B. Finite population correction factor
> #is required for frame A
>
> JackBKA (yA, yB, DatA$ProbA, DatB$ProbB, DatA$ProbB, DatB$ProbA, DatA$Domain,
+ DatB$Domain, conf_level = 0.95, sdA = "str", strA = DatA$Stratum, fcpA = TRUE)
                      Feed         Clo         Lei
Total          566434.3200 68959.26705 50953.07583
Jack Upper End 610992.1346 74715.89841 54717.32664
Jack Lower End 521876.5055 63202.63570 47188.82502
Mean              247.8845    30.17814    22.29822
Jack Upper End    267.3840    32.69738    23.94555
Jack Lower End    228.3850    27.65891    20.65090
> JackSFRR(yA, yB, DatA$ProbA, DatB$ProbB, DatA$ProbB, DatB$ProbA, DatA$Domain,
+ DatB$Domain, N_A = 1735, N_B = 1191, conf_level = 0.95, sdA = "str",
+ strA = DatA$Stratum, fcpA = TRUE)
                      Feed         Clo         Lei
Total          584713.4070 71086.18669 52423.74035
Jack Upper End 619959.0338 76576.74587 55204.67760
Jack Lower End 549467.7802 65595.62751 49642.80309
Mean              248.2219    30.17743    22.25487
Jack Upper End    263.1843    32.50828    23.43543
Jack Lower End    233.2595    27.84659    21.07431
> JackCalSF(yA, yB, DatA$ProbA, DatB$ProbB, DatA$ProbB, DatB$ProbA, DatA$Domain,
+ DatB$Domain, N_A = 1735, N_B = 1191, N_ab = 601, conf_level = 0.95, sdA = "str",
```

```
+ strA = DatA$Stratum, fcpA = TRUE)
                      Feed         Clo         Lei
Total          577163.6066 70173.20412 51726.19862
Jack Upper End 599105.4275 73516.53187 53165.97439
Jack Lower End 555221.7858 66829.87636 50286.42285
Mean              248.2424    30.18202    22.24783
Jack Upper End    257.6798    31.62001    22.86709
Jack Lower End    238.8051    28.74403    21.62857
>
> #Same for a selection of dual frame estimators
> JackHartley (yA, yB, DatA$ProbA, DatB$ProbB, DatA$Domain, DatB$Domain,
+ conf_level = 0.95, sdA = "str", strA = DatA$Stratum, fcpA = TRUE)
                      Feed         Clo         Lei
Total          570867.8042 69473.86532 51284.27265
Jack Upper End 610664.7131 74907.33129 54782.33083
Jack Lower End 531070.8954 64040.39934 47786.21447
Mean              247.9484    30.17499    22.27460
Jack Upper End    265.2336    32.53494    23.79393
Jack Lower End    230.6631    27.81504    20.75527
> JackPML(yA, yB, DatA$ProbA, DatB$ProbB, DatA$Domain, DatB$Domain,
+ N_A = 1735, N_B = 1191, conf_level = 0.95, sdA = "str", strA = DatA$Stratum,
+ fcpA = TRUE)
                      Feed         Clo         Lei
Total          594400.6320 72430.05834 53408.30337
Jack Upper End 626443.7529 76885.06491 56003.77592
Jack Lower End 562357.5111 67975.05176 50812.83082
Mean              248.0934    30.23115    22.29178
Jack Upper End    261.4677    32.09060    23.37509
Jack Lower End    234.7191    28.37171    21.20847
> JackCalDF(yA, yB, DatA$ProbA, DatB$ProbB, DatA$Domain, DatB$Domain, N_A = 1735,
```

```
+ N_B = 1191, N_ab = 601, conf_level = 0.95, sdA = "str", strA = DatA$Stratum,
+ fcpA = TRUE)
                        Feed         Clo         Lei
Total           578895.6961 70230.11306 51570.55683
Jack Upper End  601626.7000 73614.66702 53037.42260
Jack Lower End  556164.6921 66845.55910 50103.69107
Mean               248.9874    30.20650    22.18088
Jack Upper End     258.7642    31.66222    22.81179
Jack Lower End     239.2106    28.75078    21.54997
```

A2.4 An application to a real telephone survey

In the example above data are separated into two data sets `DatA` and `DatB` containing domain information. But in practice, it is common to have a joint data set including units from both samples in which there is not a specific variable indicating the domain where each individual is placed. However, we can easily split the dataset and format it, so functions of *Frames2* can be applied. To illustrate how to do this, we are going to use dataset `Dat`, which includes some of the variables collected in a real dual frame survey.

Data included in `Dat` comes from an opinion survey on the Andalusian population with respect to immigration. This survey is conducted using telephone interviews on adults using two sampling frames: one for landlines and another one for cell phones. From the landline frame, a stratified sample of size 1919 was drawn, while from the cell phone frame, a sample of size 483 is drawn using simple random sampling without replacement. First-order inclusion probabilities were computed from a stratified random design in the landline frame and modified taking into account the number of fixed lines and adults in the household. In the cell phone frame first-order inclusion probabilities were computed and modified, given the number of cell phone numbers per individual. At the time of data collection, frame sizes of land and cell phones were 4,982,920 and 5,707,655, respectively, and the total population size was 6,350,916.

The data set includes information about 7 variables: `Drawnby`, which takes value 1 if the unit comes from the landline sample and value 2 if it comes from the cell phone sample; `Stratum`, which indicates the stratum each unit belongs to (for individuals in cell phone frame, value of this variable is `NA`); `Opinion` the response to the question: "Do you think that immigrants currently living in Andalusia are quite a

lot?" with value 1 representing "yes" and value 0 representing "no"; `Landline` and `Cell`, which record whether the unit possess a landline or a cell phone, respectively. First order inclusion probabilities are also included in the data set.

```
> data(Dat)
> head(Dat,3)
  Drawnby Stratum Opinion Landline Cell ProbLandline ProbCell
1       1       2       0        1    1  0.000673623 8.49e-05
2       1       5       1        1    1  0.002193297 5.86e-05
3       1       1       0        1    1  0.001831489 7.81e-05
```

From the data of this survey we wish to estimate the number of people in Andalusia thinking that immigrants currently living in this region are quite a lot. In order to use functions of *Frames2*, we need to split this dataset. The variables we will use to do this are `Drawnby` and `Landline` and `Cell`.

```
> attach(Dat)
> #We can split the original dataset in four new different
> #datasets, each one corresponding to one domain.
>
> DomainOnlyLandline <- Dat[Landline == 1 & Cell == 0,]
> DomainBothLandline <- Dat[Drawnby == 1 & Landline == 1 & Cell == 1,]
> DomainOnlyCell <- Dat[Landline == 0 & Cell == 1,]
> DomainBothCell <- Dat[Drawnby == 2 & Landline == 1 & Cell == 1,]
>
> #From the domain datasets, we can build frame datasets
>
> FrameLandline <- rbind(DomainOnlyLandline, DomainBothLandline)
> FrameCell <- rbind(DomainOnlyCell, DomainBothCell)
>
> #Finally, we only need to label domain of each unit using "a", "b",
> #"ab" or "ba"
>
```

```
> Domain <- c(rep("a", nrow(DomainOnlyLandline)), rep("ab", nrow(DomainBothLandline)))
> FrameLandline <- cbind(FrameLandline, Domain)
>
> Domain <- c(rep("b", nrow(DomainOnlyCell)), rep("ba", nrow(DomainBothCell)))
> FrameCell <- cbind(FrameCell, Domain)
```

Now dual frame estimators, as PML estimator, can be computed:

```
> summary(PML(FrameLandline$Opinion, FrameCell$Opinion, FrameLandline$ProbLandline,
+ FrameCell$ProbCell, FrameLandline$Domain, FrameCell$Domain, N_A = 4982920,
+ N_B = 5707655))


Call:
PML(ysA = FrameLandline$Opinion, ysB = FrameCell$Opinion,
    pi_A = FrameLandline$ProbLandline, pi_B = FrameCell$ProbCell,
    domains_A = FrameLandline$Domain, domains_B = FrameCell$Domain,
    N_A = 4982920, N_B = 5707655)


Estimation:
              [,1]
Total 3.231325e+06
Mean  4.635634e-01


Variance Estimation:
                   [,1]
Var. Total 1.784362e+10
Var. Mean  3.672317e-04


Total Domain Estimations:
                   [,1]
Total dom. a   219145.1
```

```
Total dom. ab 2318841.9
Total dom. b  1346646.1
Total dom. ba 1457501.0

Mean Domain Estimations:
                  [,1]
Mean dom. a  0.4438149
Mean dom. ab 0.4990548
Mean dom. b  0.4172797
Mean dom. ba 0.4674919

Parameters:

gamma 0.3211534
```

As overlap domain size is known, we can include additionally this information in the process and compute more accurate estimators as CalDF and CalSF.

```
> summary(CalDF(FrameLandline$Opinion, FrameCell$Opinion, FrameLandline$ProbLandline,
+ FrameCell$ProbCell, FrameLandline$Domain, FrameCell$Domain, N_A = 4982920,
+ N_B = 5707655, N_ab = 4339659))

Call:
CalDF(ysA = FrameLandline$Opinion, ysB = FrameCell$Opinion,
    pi_A = FrameLandline$ProbLandline, pi_B = FrameCell$ProbCell,
    domains_A = FrameLandline$Domain, domains_B = FrameCell$Domain,
    N_A = 4982920, N_B = 5707655, N_ab = 4339659)

Estimation:
              [,1]
Total 2.985028e+06
```

```
Mean  4.700153e-01

Variance Estimation:
                   [,1]
Var. Total 1.478990e+10
Var. Mean  3.666844e-04

Parameters:

eta 0.7296841
>
> summary(CalSF(FrameLandline$Opinion, FrameCell$Opinion, FrameLandline$ProbLandline,
+ FrameCell$ProbCell, FrameLandline$ProbCell, FrameCell$ProbLandline,
+ FrameLandline$Domain, FrameCell$Domain, N_A = 4982920, N_B = 5707655, N_ab = 4339659))

Call:
CalSF(ysA = FrameLandline$Opinion, ysB = FrameCell$Opinion,
    pi_A = FrameLandline$ProbLandline, pi_B = FrameCell$ProbCell,
    pik_ab_B = FrameLandline$ProbCell, pik_ba_A = FrameCell$ProbLandline,
    domains_A = FrameLandline$Domain, domains_B = FrameCell$Domain, N_A = 4982920,
    N_B = 5707655, N_ab = 4339659)

Estimation:
              [,1]
Total 2.986787e+06
Mean  4.702923e-01

Variance Estimation:
                   [,1]
Var. Total 1.442969e+10
```

```
Var. Mean  3.577539e-04
```

Observe that as greater is the information included in the estimation process as greater is the accuracy of the estimates.

A2.5 Summary

The statistical literature about dual frame surveys started around 1960 and its development has evolved very quickly because these surveys are largely used by statistical agencies and private organizations to decrease sampling costs and to reduce frame undercoverage errors that could occur with the use of a single sampling frame.

Dual frame surveys can be more complicated to design and more complicated to analyze than those that use one frame only. There are several estimators of the population total available in the statistical literature. These estimators rely on weight adjustments to compensate the multiplicity of the units in the overlap domain. Some of these estimators allow to handle different types of auxiliary information at different levels. Nevertheless, none of the existing statistical software implements all of these estimators.

In this article we illustrate *Frames2*, a new R package for point and interval estimation in dual frame context. Functions composing the package implement the most important estimators in the literature for population totals and means. We include two procedures (Pseudo-Empirical-Likelihood approach and calibration approach) to incorporate auxiliary information about frame sizes and also about one or several auxiliary variables in one or two frames. Post-stratification, raking ratio or regression estimation are all encompassed as particular cases of these estimation procedures. Additional functions for confidence interval estimation based on the jackknife variance estimation have been included as well.

The functionalities of the package *Frames2* have been illustrated using several data sets `DatA`, `DatB` and `Dat` (included in the package) corresponding to different complex surveys. We envision future additions to the package that will allow for extensions to more than two frames.

Finally, we would like to direct the reader to the package vignettes `estimation` (*Estimation in a dual frame context*) and `formatting.data` (*Splitting and formatting data in a dual frame context*) for further examples and background information.

Appendix A3

Multinomial logistic estimation in dual frame surveys

Molina, D., Rueda, M., Arcos, A. and Ranalli, M. G. (2015)
Multinomial logistic estimation in dual frame surveys.
Statistics and Operations Research Transactions (SORT), Vol. 39, Number 2, pp. 309 - 336.

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 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.

A3.1 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, 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 (1998a) 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 A3.2, 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.

A3.2 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 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 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 A3.1. The landline sample was also stratified by provinces in the region of Andalusia, as shown in Table A3.2. 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 A3.1: 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	1727	237	1964
Mobile		246	246
Landline	192		192
Total	1919	483	2402

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,

Table A3.2: Stratification in land-phone sample

Province	Almería	Cádiz	Córdoba	Granada	Huelva	Jaén	Málaga	Sevilla
Population(*)	353787	767370	508258	558087	308941	423548	872011	1190918
Sample	262	210	252	256	275	263	207	194

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

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 A4.3.

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 A3.3: 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)

A3.3 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 $\mathcal{A}$ be the set of population units in frame A and $\mathcal{B}$ the set of population units in frame B. The population of interest, U, may be divided into three mutually exclusive domains, $a = \mathcal{A} \cap \mathcal{B}^c, b = \mathcal{A}^c \cap \mathcal{B}$ and $ab = \mathcal{A} \cap \mathcal{B}$. Because the population units in the overlap domain ab can be sampled in either survey or both surveys, it is convenient to create a duplicate domain $ba = \mathcal{B} \cap \mathcal{A}$, which is identical to $ab = \mathcal{A} \cap \mathcal{B}$, to denote the domain in the overlapping area coming from frame B. Let N, N_A, N_B, N_a, N_b, N_{ab}, N_{ba} be the number of population units in 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 $i = 1, 2, \ldots, m$. Such proportions are such that

$$P_i = N^{-1}(Z_{ai} + \eta Z_{abi} + (1 - \eta) Z_{bai} + Z_{bi}), \tag{A3.1}$$

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

$$\hat{P}_{Hi}(\eta) = N^{-1}(\hat{Z}_{ai} + \eta\hat{Z}_{abi} + (1-\eta)\hat{Z}_{bai} + \hat{Z}_{bi}), \tag{A3.2}$$

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

$$d_k^{\circ} = \begin{cases} d_{Ak} & if\ k \in s_a \\ \eta d_{Ak} & if\ k \in s_{ab} \\ (1-\eta)d_{Bk} & if\ k \in s_{ba} \\ d_{Bk} & if\ k \in s_b \end{cases}, \tag{A3.3}$$

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:

$$\hat{P}_{FBi}(\beta_1, \beta_2) = N^{-1}(\hat{Z}_{ai} + \beta_1\hat{Z}_{abi} + (1-\beta_1)\hat{Z}_{bai} + \hat{Z}_{bi} + \beta_2(\hat{N}_{ab} - \hat{N}_{ba})) \tag{A3.4}$$

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:

$$\hat{P}_{PMLi}(\theta) = N^{-1}\left(\frac{N_A - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_a}\hat{Z}_{ai} + \frac{\hat{N}_{ab}^{PML}(\theta)}{\hat{N}_{ab}(\theta)}\hat{Z}_{abi}(\theta) + \frac{N_B - \hat{N}_{ab}^{PML}(\theta)}{\hat{N}_b}\hat{Z}_{bi}\right) \tag{A3.5}$$

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 (2006) from one-frame surveys to dual-frame surveys following a stratification approach. In particular,

$$\hat{P}_{PELi}(\theta) = (N_a/N)\hat{\bar{Z}}_{aip} + \theta(N_{ab}/N)\hat{\bar{Z}}_{abip} + (1-\theta)(N_{ab}/N\hat{\bar{Z}}_{baip} + (N_b/N)\hat{\bar{Z}}_{bip}, \tag{A3.6}$$

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 (A3.6) can be rewritten as:

$$\hat{P}_{PELi} = (N_a/N)\hat{\bar{Z}}_{aip} + (N_{ab}/N)\hat{\bar{Z}}_{abip} + (N_b/N)\hat{\bar{Z}}_{bip}, \tag{A3.7}$$

so the estimator does not depend on explicitly on θ and its value only affects the estimator $\hat{\bar{Z}}_{aip}$ 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,

$$\hat{P}_{CalDFi} = N^{-1}(\sum_{k\in s} d_k^{DF} z_{ki}) \tag{A3.8}$$

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

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

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 (Banker (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,

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

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

$$\hat{P}_{BKAi} = N^{-1}\left(\sum_{k\in s_A} \tilde{d}_k z_{ki} + \sum_{k\in s_B} \tilde{d}_k z_{ki}\right) = N^{-1}\left(\sum_{k\in s} \tilde{d}_k z_{ki}\right). \tag{A3.9}$$

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:

$$\hat{P}_{CalSFi} = N^{-1}\left(\sum_{k\in s} d_k^{SF} z_{ki}\right) \tag{A3.10}$$

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)).

A3.4 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 $\mathbf{y} = (y_1, ..., y_N)^T$ is the determination of a set of superpopulation random variables $\mathbf{Y} = (Y_1, ..., Y_N)^T$ s.t.

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

that is, we use the multinomial logistic model to relate y and $\boldsymbol{x}$. Let $\boldsymbol{\beta}$ be the parameter vector $(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_m^T)^T$. 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.

A3.4.1 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

$$\ell_{d^\circ}(\boldsymbol{\beta}) = \sum_{i=1,...,m}\left(\sum_{k\in s_A} d_k^\circ z_{ki}\ln\mu_{ki} + \sum_{k\in s_B} d_k^\circ z_{ki}\ln\mu_{ki}\right). \tag{A3.11}$$

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,

$\widehat{\boldsymbol{\beta}}^o$ obtained using the the π-weighted likelihood (A3.11) is its design based estimate. Computing $\widehat{\boldsymbol{\beta}}^o$ 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 $\widehat{\boldsymbol{\beta}}^o$ of $\boldsymbol{\beta}$, we consider the following auxiliary variable

$$p_{ki}^o = \widehat{\mu}_{ki}^o = \frac{\exp(\boldsymbol{x}_k^T \widehat{\boldsymbol{\beta}}_i^o)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_k^T \widehat{\boldsymbol{\beta}}_r^o)}. \tag{A3.12}$$

Please note that these p values are different from those involved in the definition of estimator (A3.6). Since the vector $\boldsymbol{x}_k$ is known for all units of the population U, the values p_{ki}^o are available for all $k \in U$ and we propose to use such values to define a new estimator for P_i,

$$\begin{aligned} \widehat{P}_{MLi}^{DW} &= N^{-1}\left(\sum_{k\in U} p_{ki}^o + \sum_{k\in s_A} d_k^\circ (z_{ki} - p_{ki}^o) + \sum_{k\in s_B} d_k^\circ (z_{ki} - p_{ki}^o) \right) \\ &= N^{-1}\left(\sum_{k\in U} p_{ki}^o + \sum_{k\in s} d_k^\circ (z_{ki} - p_{ki}^o) \right). \end{aligned} \tag{A3.13}$$

We observe that this estimator takes the same model-assisted form as the MLGREG estimator proposed in Lehtonen and Veijanen (1998a), 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 $\sum_{k\in U} p_{ki}^o$ in (A3.13) 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 A4.3.

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

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

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:

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

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

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

$$\ell_{\tilde{d}}(\boldsymbol{\beta}) = \sum_{i=1,\ldots,m} \sum_{k\in s} \tilde{d}_k z_{ki} \ln \mu_{ki}. \tag{A3.14}$$

We use the new auxiliary variable $\tilde{p}_{ki} = \tilde{\mu}_{ki} = \dfrac{\exp(\boldsymbol{x}_k^T \tilde{\boldsymbol{\beta}}_i)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_k^T \tilde{\boldsymbol{\beta}}_r)}$ 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):

$$\begin{aligned} \widehat{P}^{SW}_{MLi} &= N^{-1}\left(\sum_{k\in U} \tilde{p}_{ki} + \sum_{k\in s_A} \tilde{d}_k(z_{ki}-\tilde{p}_{ki}) + \sum_{k\in s_B} \tilde{d}_k(z_{ki}-\tilde{p}_{ki})\right) \\ &= N^{-1}\left(\sum_{k\in U} \tilde{p}_{ki} + \sum_{k\in s} \tilde{d}_k(z_{ki}-\tilde{p}_{ki})\right). \end{aligned} \tag{A3.15}$$

Note that $\tilde{d}_k$ weights are used in the formulation of the estimator (A3.15) and also in the likelihood

function (A3.14).

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):

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

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.

A3.4.2 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 $k \in \mathcal{A}$ we have an auxiliary vector $\boldsymbol{x}_{Ak}$ and for each unit $k \in \mathcal{B}$ 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 $k \in \mathcal{A}$, using data from s_A we can compute

$$p^A_{ki} = \frac{\exp(\boldsymbol{x}^T_{Ak}\widehat{\boldsymbol{\beta}}^A_i)}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}^T_{Ak}\widehat{\boldsymbol{\beta}}^A_r)} \tag{A3.16}$$

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

$$\widehat{P}_{MLi}^{DF} = N^{-1} \left(\sum_a p_{ki}^A + \eta \sum_{ab} p_{ki}^A + (1-\eta) \sum_{ba} p_{ki}^B + \sum_b p_{ki}^B + \right.$$

$$+ \sum_{s_a} (z_{ki} - p_{ki}^A) d_{Ak} + \eta \sum_{s_{ab}} (z_{ki} - p_{ki}^A) d_{Ak} +$$

$$\left. + (1-\eta) \sum_{s_{ba}} (z_{ki} - p_{ki}^B) d_{Bk} + \sum_{s_b} (z_{ki} - p_{ki}^B) d_{Bk} \right).$$

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 $\widehat{P}_{MLi}^{DF}$ 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,

$$\widehat{P}_{MLCi}^{DF} = N^{-1} (\sum_{k \in s_A} w_k^\star z_{ki} + \sum_{k \in s_B} w_k^\star z_{ki}) = N^{-1} (\sum_{k \in s} w_k^\star z_{ki}), \tag{A3.17}$$

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

$$\min \sum_{k \in s_A} G(w_k^\star, d_{Ak}) + \sum_{k \in s_B} G(w_k^\star, d_{Bk}) \qquad \text{s.t.}$$

$$\sum_{k \in s_A} w_k^\star p_{ki}^A = \sum_{k \in a} p_{ki}^A + \eta \sum_{k \in ab} p_{ki}^A,$$

$$\sum_{k \in s_B} w_k^\star p_{ki}^B = (1-\eta) \sum_{k \in ba} p_{ki}^B + \sum_{k \in b} p_{ki}^B,$$

$$\sum_{k \in s_a} w_k^\star = N_a, \quad \sum_{k \in s_b} w_k^\star = N_b,$$

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

where p_{ki}^A are the estimated probabilities defined in (A3.16) and p_{ki}^B are their analogous in frame B.

A3.5 Properties of proposed estimators

To show the asymptotic properties of the proposed estimators $\hat{P}_{ML}^{DW}$, $\hat{P}_{MLC}^{DW}$, $\hat{P}_{ML}^{SW}$, $\hat{P}_{MLC}^{SW}$, $\hat{P}_{ML}^{DF}$, $\hat{P}_{MLC}^{DF}$, 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 $\hat{P}_{MLC}^{DW}$ and then move to the other estimators, because these can be dealt with using slight modifications of this more general setting. Let $\mu(\boldsymbol{x}_k, \boldsymbol{\theta}_i) = \exp(\boldsymbol{x}_k^T \boldsymbol{\theta}_i) / \sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_k^T \boldsymbol{\theta}_r)$, 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 $\hat{\boldsymbol{\beta}}^o = \boldsymbol{\beta}_U + O_p(n_N^{-1/2})$.*

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.*

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

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 A6 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 $\hat{P}_{MLC}^{DW}$ for the Euclidean distance. In fact, in this case an analytic solution to the constrained distance minimization problem exists and is given by

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

where $\tilde{\boldsymbol{p}}_{ki}^\circ = (\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b), p_{ki}^\circ)^T$ is a vector that contains p_{ki}° defined in (A4.15) 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, $\hat{\boldsymbol{\alpha}}_i^\circ = (\sum_{k\in s} d_k^\circ \tilde{\boldsymbol{p}}_{ki}^\circ \tilde{\boldsymbol{p}}_{ki}^{\circ T})^{-1}(\sum_{k\in s} d_k^\circ \tilde{\boldsymbol{p}}_{ki}^\circ z_{ki})$, i.e. it is the vector of coefficients of the generalized regression of z_{ki} on $\tilde{\boldsymbol{p}}_{ki}^\circ$ 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 $\widehat{P}_{MLCi}^{GDW}$, under additional regularity conditions on the shape of the distance function itself.

Theorem 1. *Under assumptions A6–A10,* $\widehat{P}_{MLCi}^{GDW}$ *is design* $\sqrt{n_N}$*-consistent for* P_i *in the sense that*

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

and has the following asymptotic distribution

$$\frac{\widehat{P}_{MLCi}^{GDW} - P_i}{\sqrt{V_\infty(\widehat{P}_{MLCi}^{GDW})}} \xrightarrow{\mathcal{L}} N(0,1)$$

where $V_\infty(\widehat{P}_{MLCi}^{GDW}) = N^{-2}V(\hat{t}_{eiH})$ *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 $\tilde{\boldsymbol{p}}_{ki}^{\circ}$ *but with* p_{ki}° *replaced by its population counterpart*

$$\mu_{ki}^{\circ} = \frac{\exp(\boldsymbol{x}_k^T \boldsymbol{\beta}_{Ui})}{\sum_{r=1,\ldots,m} \exp(\boldsymbol{x}_k^T \boldsymbol{\beta}_{Ur})}. \tag{A3.18}$$

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

$$\begin{aligned} v(\widehat{P}_{MLCi}^{GDW}) &= N^{-2} v(\hat{t}_{\hat{e}iH}) \\ &= N^{-2} \Big\{ v\Big(\textstyle\sum_{k\in s_a} d_{Ak} \hat{e}_{ki} + \eta \sum_{k\in s_{ab}} d_{Ak} \hat{e}_{ki} \Big) + \\ &\quad + v\Big((1-\eta) \textstyle\sum_{k\in s_{ba}} d_{Bk} \hat{e}_{ki} + \sum_{k\in s_b} d_{Bk} \hat{e}_{ki} \Big) \Big\}. \end{aligned} \tag{A3.19}$$

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 A6–A7 and A9–A10,

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

using a first order Taylor expansion of $\mu(\boldsymbol{x}_k, \hat{\boldsymbol{\beta}}_i^o)$ at $\hat{\boldsymbol{\beta}}_i^o = \boldsymbol{\beta}_{Ui}$, 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 A6–A10 and a second order Taylor expansion of $\mu(\boldsymbol{x}_k, \hat{\boldsymbol{\beta}}_i^o)$ at $\hat{\boldsymbol{\beta}}_i^o = \boldsymbol{\beta}_{Ui}$,

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

Then,

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

and the first part of the result is proven.

Now, from assumption A10, $v(\hat{t}_{eiH}) = V(\hat{t}_{eiH}) + o_p(n_N^{-1})$. 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 (A3.19) can be used to estimate the variance of $\hat{P}_{MLC}^{DW}$ also when using different distance functions.

Now, $\hat{P}_{ML}^{DW}$ can be seen as a particular case of $\widehat{P}_{MLCi}^{GDF}$ in which $\tilde{\boldsymbol{p}}_{ki}^{\circ}$ includes only p_{ki}°, and $\hat{\boldsymbol{\alpha}}_i^{\circ}$ is only a scalar and is set exactly equal to 1. Therefore, $\hat{P}_{ML}^{DW}$ is consistent for P_i and asymptotically normal

with $V_\infty(\hat{P}^{DW}_{ML}) = N^{-2}V(\hat{t}_{eiH})$, where census-level residuals are given here by $e_{ki} = z_{ki} - \mu^\circ_{ki}$. Variance estimation can again be conducted by plugging sample level estimated residuals in (A3.19) given in this case by $\hat{e}_{ki} = z_{ki} - p^\circ_{ki}$.

Estimator $\hat{P}^{DF}_{MLC}$ is in all similar to $\hat{P}^{DW}_{MLC}$, 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 $\tilde{\boldsymbol{p}}^{A,B}_{ki}$ and contains p^A_{ki}, p^B_{ki} and the other indicator variables used in the benchmark constraints: for example $\tilde{\boldsymbol{p}}^{A,B}_{ki} = (\delta_k(a), \delta_k(ab), \delta_k(ba), \delta_k(b), [\delta_k(a) + \delta_k(ab)]p^A_{ki}, [\delta_k(b) + \delta_k(ba)]p^B_{ki})^T$.

To encompass this situation, it is enough to change assumption A6 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 $\hat{P}^{DF}_{MLC}$ 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, $V_\infty(\widehat{P}^{GDF}_{MLCi}) = N^{-2}V(\hat{t}_{eiH})$ and $\hat{t}_{eiH} = \sum_{k\in s} d^\circ_k e_{ki}$ is the Hartley estimator of the population total of the census-level residuals given here by $e_{ki} = z_{ki} - (\tilde{\boldsymbol{\mu}}^{A,B})^T_{ki}\boldsymbol{\alpha}_i$, where $\tilde{\boldsymbol{\mu}}^{A,B}_{ki}$ is like $\tilde{\boldsymbol{p}}^{A,B}_{ki}$ but with p^A_{ki} and p^B_{ki} replaced by their population counterparts, similarly to (A3.18). Analytic variance estimation can be conducted by using sample level estimates of the residuals. In particular, by using $\hat{e}_{ki} = z_{ki} - (\tilde{\boldsymbol{p}}^{A,B}_{ki})^T\hat{\boldsymbol{\alpha}}_i$ in formula (A3.19).

Now, similarly as for $\hat{P}^{DW}_{ML}$ and $\hat{P}^{DW}_{MLC}$, $\hat{P}^{DF}_{ML}$ can be seen as a particular case of $\widehat{P}^{GDF}_{MLCi}$ in which $\tilde{\boldsymbol{p}}^\circ_{ki}$ includes only $p^{A,B}_{ki}$, with $p^{A,B}_{ki} = p^A_{ki}$ if $k \in s_A$ and $p^{A,B}_{ki} = p^B_{ki}$ if $k \in s_B$, and $\hat{\boldsymbol{\alpha}}^\circ_i$ 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 $V_\infty(\hat{P}^{DF}_{ML}) = N^{-2}V(\hat{t}_{eiH})$, where census-level residuals are given here by $e_{ki} = z_{ki} - \mu^{A,B}_{ki}$, and $\mu^{A,B}_{ki}$ is the census level fit corresponding to $p^{A,B}_{ki}$. Variance estimation can again be conducted by using sample level estimated residuals in equation (A3.19) given by $\hat{e}_{ki} = z_{ki} - p^A_{ki}$ if $k \in s_A$ and $\hat{e}_{ki} = z_{ki} - p^B_{ki}$ if $k \in s_B$.

The calibration estimator $\hat{P}^{SW}_{MLC}$ is very similar to $\hat{P}^{DW}_{MLC}$, the only differences are (i) in the set of basic design weights employed in the calibration procedure: for $\hat{P}^{SW}_{MLC}$ we use $\tilde{d}_k$, and (ii) p°_{ki} is replaced by $\tilde{p}_{ki}$ in the definition of the vector $\tilde{\boldsymbol{p}}^\circ_{ki}$. Once these changes are incorporated across assumption A6, and assumption A10 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 $\hat{P}^{SW}_{MLC}$ is given by $V_\infty(\widehat{P}^{GSW}_{MLCi}) = N^{-2}V(\hat{t}_{ei})$ 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 $e_{ki} = z_{ki} - \tilde{\boldsymbol{\mu}}_{ki}^T \boldsymbol{\alpha}_i$, and where $\tilde{\boldsymbol{\mu}}_{ki}$ is like $\tilde{\boldsymbol{p}}_{ki}$ but with p_{ki} replaced by its population counterpart

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

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

A3.6 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 (2009b) for a review). Hartley (1962) proposed choosing η to minimize the variance of the estimator in (A3.2). 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 $i = 1, ..., m$ 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

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

or alternatively

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

being $V(\hat{N}_{ab}^A)$ and $V(\hat{N}_{ab}^B)$ 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 $V(\hat{N}_{ab}^A)$ and $V(\hat{N}_{ab}^B)$ 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)).

A3.7 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 (2007) 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

$$v_J(\hat{P}_i) = V_J^A + V_J^B = \frac{n_A - 1}{n_A} \sum_{g \in s_A} (\hat{P}_i^A(g) - \overline{P}_i^A)^2 + \frac{n_B - 1}{n_B} \sum_{j \in s_B} (\hat{P}_i^B(j) - \overline{P}_i^B)^2 \tag{A3.20}$$

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

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

$$= \sum_{h=1}^{H} \frac{n_{Ah} - 1}{n_{Ah}} \sum_{g \in s_{Ah}} (\hat{P}_i^A(hg) - \overline{P}_i^{Ah})^2 + \sum_{l=1}^{L} \frac{n_{Bl} - 1}{n_{Bl}} \sum_{j \in s_{Bl}} (\hat{P}_i^B(lj) - \overline{P}_i^{Bl})^2, \tag{A3.21}$$

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:

$$v_{Jc}(\hat{P}_i) = \frac{n_A - 1}{n_A}(1 - \overline{\pi}_A) \sum_{g \in s_A} (\hat{P}_i^A(g) - \overline{P}_i^A)^2 + \frac{n_B - 1}{n_B}(1 - \overline{\pi}_B) \sum_{j \in s_B} (\hat{P}_i^B(j) - \overline{P}_i^B)^2 \tag{A3.22}$$

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

$$v_{Jc}^{st}(\hat{P}_i) = \sum_{h=1}^{H} \frac{n_{Ah} - 1}{n_{Ah}}(1 - \overline{\pi}_{Ah}) \sum_{g \in s_{Ah}} (\hat{P}_i^A(hg) - \overline{P}_i^{Ah})^2$$

$$+ \sum_{l=1}^{L} \frac{n_{Bl} - 1}{n_{Bl}}(1 - \overline{\pi}_{Bl}) \sum_{j \in s_{Bl}} (\hat{P}_i^B(lj) - \overline{P}_i^{Bl})^2 \tag{A3.23}$$

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 simple work with the ultimate cluster rather than elementary units (see e.g. Wolter (2007)).

A3.8 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 $N = 10000$.

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}=1000 and the resulting sizes of the two frames are N_A=6000 and N_B=5000. 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=6000 and N_B=7000 and the overlap domain size is N_{ab}=3000, while for the third scenario we have N_A=8000, N_B=7000 and N_{ab}=5000. 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 $\hat{P}^{DW}_{ML}$, $\hat{P}^{DW}_{MLC}$, $\hat{P}^{SW}_{ML}$ and $\hat{P}^{SW}_{MLC}$ estimators using as x-variable `ses` (Case I), and $\hat{P}^{DF}_{ML}$ and $\hat{P}^{DF}_{MLC}$ 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 (A3.9), 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 ($\hat{P}_{CalDF}$ and $\hat{P}_{CalSF}$) 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 $\hat{P}_{CalDF}$ estimator and as x-variable `ses` for $\hat{P}_{CalSF}$ 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 $RB\% = E_{MC}(\hat{Y} - Y)/Y * 100$, the percent relative mean squared error $RMSE\% = E_{MC}[(\hat{Y} - Y)^2]/Y^2 * 100$, 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 A3.4 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 ($\hat{P}^{DW}_{ML}$, $\hat{P}^{DW}_{MLC}$, $\hat{P}^{SW}_{ML}$, $\hat{P}^{DW}_{MLC}$, $\hat{P}^{DF}_{ML}$ and $\hat{P}^{DF}_{MLC}$).

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 A3.5 shows the average length reduction of 95% confidence intervals and 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, $\hat{P}^{DW}_{ML}$, $\hat{P}^{DW}_{MLC}$, $\hat{P}^{SW}_{ML}$ 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 $\hat{P}_{CalDF}$.

From Table A3.5 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 $\boldsymbol{\beta}$ ($\hat{P}_{ML}^{DW}$, $\hat{P}_{MLC}^{DW}$, $\hat{P}_{ML}^{SW}$ and $\hat{P}_{MLC}^{SW}$) 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.

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

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

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

See Section A3.6 for details and guidelines on choosing a value for η. Table A3.6 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.

A3.9 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 A4.3. 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 A3.7 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. Calibration approach achieves most important reductions in length, with single frame calibration presenting the best results. On the other hand, using $\hat{P}_{ML}^{DW}$, $\hat{P}_{ML}^{SW}$ and $\hat{P}_{ML}^{DF}$ estimators the length reduction is less noticeable.

Table A3.8 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.

A3.10 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 A3.9.

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.

Table A3.4: Relative efficiency (respect to the BKA estimator) of compared estimators. POP1 and POP2

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

Table A3.5: 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

Table A3.6: 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.
$\hat{P}^{DW}_{ML}$	$\hat{\eta}_{SR2}$	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
$\hat{P}^{DW}_{MLC}$	$\hat{\eta}_{SR2}$	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
$\hat{P}^{DF}_{ML}$	$\hat{\eta}_{SR2}$	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
$\hat{P}^{DF}_{MLC}$	$\hat{\eta}_{SR2}$	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

Table A3.7: 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				
$\hat{P}_{ML}^{DW}$	42.75	39.76	45.74	5.98	14.33
$\hat{P}_{MLC}^{DW}$	41.23	38.78	43.68	4.90	29.80
$\hat{P}_{ML}^{SW}$	42.89	39.94	45.84	5.90	15.47
$\hat{P}_{MLC}^{SW}$	41.41	39.03	43.79	4.76	31.81
$\hat{P}_{ML}^{DF}$	42.61	39.64	45.58	5.94	14.90
$\hat{P}_{MLC}^{DF}$	41.16	38.67	43.65	4.98	28.65
	A reasonable number				
$\hat{P}_{ML}^{DW}$	45.24	42.27	48.20	5.93	12.28
$\hat{P}_{MLC}^{DW}$	46.57	44.11	49.03	4.92	27.22
$\hat{P}_{ML}^{SW}$	45.09	42.17	48.01	5.84	13.61
$\hat{P}_{MLC}^{SW}$	46.40	44.02	48.78	4.76	29.59
$\hat{P}_{ML}^{DF}$	45.45	42.49	48.41	5.92	12.43
$\hat{P}_{MLC}^{DF}$	46.68	44.17	49.18	5.01	25.89
	Too few				
$\hat{P}_{ML}^{DW}$	6.06	4.55	7.58	3.03	15.36
$\hat{P}_{MLC}^{DW}$	5.77	4.58	6.97	2.39	33.24
$\hat{P}_{ML}^{SW}$	6.05	4.56	7.54	2.98	16.76
$\hat{P}_{MLC}^{SW}$	5.76	4.61	6.91	2.30	35.75
$\hat{P}_{ML}^{DF}$	6.13	4.62	7.64	3.02	15.64
$\hat{P}_{MLC}^{DF}$	5.63	4.46	6.80	2.34	34.64
	No reply				
$\hat{P}_{ML}^{DW}$	5.95	4.65	7.25	2.60	12.75
$\hat{P}_{MLC}^{DW}$	6.43	5.27	7.58	2.31	22.48
$\hat{P}_{ML}^{SW}$	5.96	4.67	7.25	2.58	13.42
$\hat{P}_{MLC}^{SW}$	6.43	5.30	7.56	2.26	24.16
$\hat{P}_{ML}^{DF}$	5.80	4.51	7.10	2.59	13.09
$\hat{P}_{MLC}^{DF}$	6.54	5.33	7.74	2.41	19.13

Table A3.8: 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				
$\hat{P}_{ML}^{DW}$	42.75	46.46	39.15	32.46	44.29	46.03	45.14
$\hat{P}_{MLC}^{DW}$	41.23	43.64	38.97	30.97	42.07	43.31	46.58
$\hat{P}_{ML}^{SW}$	42.89	46.74	39.11	32.76	43.89	46.44	45.85
$\hat{P}_{MLC}^{SW}$	41.41	43.79	39.19	31.55	41.61	43.87	45.77
$\hat{P}_{ML}^{DF}$	42.61	44.45	39.16	31.99	41.69	43.56	48.13
$\hat{P}_{MLC}^{DF}$	41.16	43.55	38.96	30.01	42.14	43.28	48.56
			A reasonable number				
$\hat{P}_{ML}^{DW}$	45.24	42.31	48.10	59.82	40.71	40.72	44.47
$\hat{P}_{MLC}^{DW}$	46.57	44.39	48.74	61.97	44.44	42.72	43.25
$\hat{P}_{ML}^{SW}$	45.09	42.04	48.11	59.62	40.90	40.68	43.70
$\hat{P}_{MLC}^{SW}$	46.40	44.14	48.63	61.49	44.67	42.64	43.61
$\hat{P}_{ML}^{DF}$	45.45	44.02	48.35	60.42	43.98	42.81	42.11
$\hat{P}_{MLC}^{DF}$	46.68	44.59	48.78	63.21	44.46	42.56	41.65
			Too few				
$\hat{P}_{ML}^{DW}$	6.06	6.75	5.35	3.77	9.84	6.18	2.82
$\hat{P}_{MLC}^{DW}$	5.77	6.68	4.92	3.29	7.58	6.73	2.80
$\hat{P}_{ML}^{SW}$	6.05	6.64	5.47	3.79	9.89	6.12	2.83
$\hat{P}_{MLC}^{SW}$	5.76	6.67	4.92	3.39	7.62	6.66	2.95
$\hat{P}_{ML}^{DF}$	6.13	6.58	5.11	3.50	8.17	6.37	2.39
$\hat{P}_{MLC}^{DF}$	5.63	6.46	4.81	2.92	7.46	6.77	2.35
			No reply				
$\hat{P}_{ML}^{DW}$	5.95	4.47	7.39	3.95	5.16	7.06	7.56
$\hat{P}_{MLC}^{DW}$	6.43	5.28	7.37	3.76	5.91	7.24	7.37
$\hat{P}_{ML}^{SW}$	5.96	4.58	7.31	3.83	5.32	6.76	7.62
$\hat{P}_{MLC}^{SW}$	6.43	5.41	7.26	3.57	6.10	6.84	7.67
$\hat{P}_{ML}^{DF}$	5.80	4.95	7.38	4.09	6.15	7.25	7.36
$\hat{P}_{MLC}^{DF}$	6.54	5.39	7.45	3.86	5.93	7.39	7.44

Appendix A4

Estimation of proportions for class frequencies with ordinal outcomes in multiple frame surveys with complex sampling designs

Rueda, M., Arcos, A., Molina, D. and Ranalli, M. G. (2016)
Estimation of proportions for class frequencies with ordinal outcomes in multiple frame surveys with complex sampling designs.
Survey Research Methods. In review process.

Abstract

Surveys usually include questions where individuals should select one in a series of possible options which can be somehow ordered. This kind of items are particularly frequent in social, marketing and opinion surveys where, usually, respondents are asked to indicate their degree of agreement with a list of sentences through a Likert or any other measurement scale. 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. Our aim is to estimate proportions when the variable of interest has ordinal outcomes. We propose to describe the joint distribution of the class indicators by an ordinal model. Several 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 via the empirical relative bias and the empirical relative efficiency. Empirical coverage of confidence intervals and their lengths are computed using jackknife techniques for variance estimation. An application to real survey data is also included.

A4.1 Introduction

Multiple 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. Since then, multiple frame sampling theory has experienced a noticeable development and several estimators for the total of a continuous variable have been proposed. First proposals were formulated in a dual frame context, i.e. for the case where two frames are available for sampling. Hartley (1962) himself proposed the first dual frame estimator, which was improved by Lund (1968) and Fuller and Burmeister (1972). Bankier (1986) and Kalton and Anderson (1986) and Skinner (1991) proposed dual frame estimators based on new techniques. Skinner and Rao (1996) and Rao and Wu (2010) applied likelihood methods to compute estimators that perform well in complex designs. More recently, Ranalli *et al.* (2015) and Elkasabi *et al.* (2015) used calibration techniques to derive estimators in the dual frame context.

In recent years, a number of works has arisen that focus on the estimation in cases with three or more sampling frames. Lohr and 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 easy to compute. Multiplicity is also used by Rao and Wu (2010) to provide an extension of the pseudo empirical likelihood estimator to the case of more than two frames. In 2011, Singh and Mecatti suggested a class of multiplicity estimators that encompasses all the multiple frames estimators available in the literature by suitably specifying a

set of parameters.

Popularity of multiple frame surveys has increased among scientific community along last years and now they are widely used both in statistical agencies and in private organizations. From 2000 to the present, there has been a steady increase in the use of telephone surveys, which have replaced all other data collection methods (the majority of which were face-to-face interviews). In some subject areas (e.g., electoral studies), face-to-face surveys have been completely ousted by telephone interviewing. Moreover, studies have reported improved results from phone surveys compared with face-to-face interviews (Abascal *et al.*, 2012). Telephone surveys also present some drawbacks with regard to coverage, due to the absence of a telephone in some households and the generalized use of mobile phones, which are sometimes replacing fixed (land) lines entirely (see Trujillo *et al.*, 2005). The potential for coverage error as a result of the exponential growth of the cell phone-only population has been a key point in the increasing of the use of a dual-frame approach when conducting telephone surveys. An example of a phone survey using both a landline and a cell frame to ensure the highest possible coverage of the eligible population is the 2014 U.S. National Survey of Latinos (Lopez *et al.*, 2014). Jackson *et al.* (2014) and McMillen *et al.* (2015) compared the estimates obtained through a dual frame survey with those computed using a single frame survey, with similar results. Surveys where data are collected from three sampling frames are also used in practice. Iachan *et al.* (1993) used a three frame survey to reach the homeless population of Washington D.C. metropolitan area. Frames in this survey were composed of homeless shelters, soup kitchens and street areas. On the other hand, the Canadian Community Health Survey conducted by Statistics Canada (2003) is based on a area frame, a list frame and a RDD frame.

The internet has become a very important data source that offers inexpensive ways to collect information. Couper (2000) analyzes the issues and challenges related with web surveys concluding that this kind of surveys already offer enormous potential for survey researchers which is likely only to improve with time. Within multiple frame context, Lohr (2010) points that web surveys will play a very important role in the future development of multiple frame surveys. So, in the near future it is very likely that dual frame surveys consisting of a cell and a landline frame evolve to multiple frame surveys incorporating a third frame of web users.

Surveys in general, and multiple frame surveys in particular, 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, eparticularly, 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 somehow ordered according to a particular criteria (e.g., from the worst to the best opinion). The main aim is to estimate the proportion of individuals selecting each option. Although classical multiple frame estimators can be used, the estimates they provide are inconsistent since they do not add up to 1 through all the categories. In the dual frame setting, Molina *et al.* (2015) developed some logistic multinomial estimators but they may not be a good choice either, since for ordinal variables the distance between adjacent categories is unknown and cannot be assumed as equal, which is a basic assumption in the multinomial approach.

In this context, the ordinal logit model (OLM) offers a great power for the estimation. The most popular OLM is the cumulative logit model, where categories of the variable of interest are divided into two groups: the first one containing a particular category together with all categories lying below, and the second one including categories above that particular category. When working with cumulative logit models it is common to assume the proportional odds (PO) property, which establishes that the distance between categories, even though unknown, is equivalent. That is, for each predictor variable the estimated cumulative odds of being at or below a particular level of the response variable are assumed to be the same across all the ordinal categories. Assuming this property leads to a more parsimonious model and, consequently, to simpler interpretations. The cumulative OLM with PO property is considered as the default ordinal regression model in the most common used statistical softwares, such as SPSS, SAS or Stata.

Although ordinal regression models have been extensively used in sociological, medical and educational applications, its use for parameter estimation in finite populations sampling is very sparse.

This article proceeds as follows: Section 2 introduces the problem of estimating the proportions of an ordinal response variable in a multiple frame context, reviewing the existing approaches for estimation. In section 3, we propose some estimators based on the ordinal logistic regression for estimating proportions of a response with ordinal outcomes using model assisted and model calibrated techniques. Main theoretical properties of the proposed estimators are studied in section 4. Performance of the estimators will be measured through simulation experiments in section 5. Finally, we check how the estimators work in a real context by applying them to data corresponding to a survey on perceptions of immigration in a certain region in section 6.

A4.2 Existing approaches for estimating proportions of a variable with ordinal outcomes in a multiple frame context

We will employ the notation used 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 $N_1, \ldots, N_q, \ldots, N_Q$, all of them can be incomplete but it is assumed that overall they cover the entire target population U. 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. 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 a discrete m-valued survey variable y and we denote 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 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 population proportion for each i, that is

$$P_i = \frac{1}{N} \sum_{k \in U} z_{ki}, \quad i = 1, 2, \ldots, m. \tag{A4.1}$$

Note that these proportions can be rewritten as follows

$$P_i = \frac{1}{N} \sum_{q=1}^{Q} \sum_{k \in U_q} \frac{z_{ki}}{m_k}, \quad i = 1, 2, \ldots, m, \tag{A4.2}$$

where m_k indicates the number of frames unit k belongs to, i.e. the multiplicity of k.

Let s_q be a sample drawn from frame A_q under a particular sampling design $p_q(s_q)$, 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 units in frame q. Let n_q be the size of sample s_q and that $s = \cup_q s_q$. For ease of notation, we will drop (q) from probabilities and weights, i.e. we will consider $\pi_k = \pi_k(q)$, $\pi_{kl} = \pi_{kl}(q)$ and $d_k = d_k(q)$, when this is not ambiguous.

Lohr and Rao (2006) formulated the multiple frame extension of some of the estimators originally proposed for the dual frame case, as the one proposed by Hartley (1962, 1974) or by Fuller and Burmeister

(1972). Although the optimal version of these estimators is asymptotically efficient, it is not internally consistent since a different set of weights is sued for each response variable. Moreover, it is often unstable in small or moderate samples with more than two frames because the optimal estimated parameters involved in the computation of the estimators are functions of large estimated covariances matrices. Lohr and Rao (2006) also followed the so called single frame approach used by Kalton and Anderson to proposed a single frame estimator in a multiple frame context. This estimator is in the form:

$$\hat{P}_{KAi} = \frac{1}{N} \sum_{k \in s} z_{ki} d_k^{KA} \tag{A4.3}$$

with $d_k^{KA} = \bar{\pi}_k^{-1}$, where $\bar{\pi}_k = \sum_{q' \ni k} \pi_k(q')$. 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. This can be an important drawback particularly if misclassification issues are present.

Lohr and Rao also proposed the following pseudo-maximum likelihood estimator for the multiple frame context:

$$\hat{P}_{PMLi} = \frac{1}{N} \sum_{k \in s} z_{ki} d_k^{PML}(q), \tag{A4.4}$$

where the weights d_k^{PML} can be defined as

$$d_k^{PML}(q) = d_k(q) f(q) \sum_{K: q \in K} \frac{\hat{N}_K \delta_k(K)}{\sum_{j \in K} f(j) \hat{N}_K(j)}$$

with $f(q) = \frac{1}{deff_z(q)} \frac{n_q}{N_q}$, being $deff_z(q)$ the design effect for variable z in the q-th frame. Values $\hat{N}_K(q)$ can be computed as $\hat{N}_K(q) = \sum_{k \in s_q} d_k(q) \delta_k(K)$, with $\delta_k(K)$ the indicator variable for domain K that takes the value 1 whether unit k belongs to domain K and 0 otherwise. The estimated domain sizes $\hat{N}_K$ are the solution of a system of non linear equations. The pseudo maximum likelihood is consistent and usually works well in practical situations but it is complex to compute for a general sampling design, since numerical procedures are required to obtain the values $\hat{N}_K$.

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

$$\hat{P}_{Mi} = \frac{1}{N} \sum_{k \in s} z_{ki} d_k^{M}, \tag{A4.5}$$

with $d_k^M = d_k/m_k$. The previous estimator, often called single frame multiplicity estimator, only requires the knowledge of the multiplicity of each unit, i.e. the number of frames the unit is included, no matter which these frames are. This estimator can be adjusted using a raking ratio approach to get a single frame raking ratio multiplicity estimator where a new set of weights, resulting from an iterative procedure, is used.

In 2011, Singh and Mecatti proposed a composite multiplicity estimator, which generalizes the single frame multiplicity estimator. This estimator can be written as

$$\hat{P}_{CMi} = \frac{1}{N}\sum_{k\in s} z_{ki} d_k^{CM} \tag{A4.6}$$

where

$$d_k^{CM} = \frac{\lambda_k d_k + (1-\lambda_k) d_k^{KA}}{m_k}$$

with

$$\lambda_k = \frac{\sum_{q'\ni k}(1-\bar{\pi}_k/\pi_k(q'))\pi_k(q')(1-\pi_k(q'))}{\sum_{q'\ni k}(1-\frac{\bar{\pi}_k^2}{\pi_k(q')^2}-\frac{2\bar{\pi}_k}{\pi_k(q')})\pi_k(q')(1-\pi_k(q'))}.$$

Usually, additional 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 the vector $\mathbf{x}_{qk} = (x_{q1k}, x_{q2k}, \ldots, x_{qp_qk})'$ includes the values of the variables $\boldsymbol{x}_q$ for the unit k of frame q. 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 coming from frame 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.

Rao and Wu (2010) followed a single frame multiplicity based approach to extend the pseudo empirical likelihood estimator for the mean of a variable to the multiple frame setting. This estimator can be computed as

$$\hat{P}_{PELi} = \sum_{k\in s} z_{ki} p_k(q) \tag{A4.7}$$

with $p_k(q)$ maximizing the likelihood function

$$l_{PEL}(\mathbf{p}_1, \ldots, \mathbf{p}_Q) = \frac{\sum_{q=1}^{Q} n_q}{\sum_{k\in s} d_k^M} \sum_{k\in s} d_k^M \log[p_k(q)]$$

subject to

$$\sum_{k\in s} p_k(q) = 1$$

$$\sum_{k\in s} p_k(q)\mathbf{x}_k = \bar{\boldsymbol{x}}$$

being $\bar{\boldsymbol{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)'$ the vector of the population means of variables $\boldsymbol{x}_q$, which are assumed in this case to be the same in all frames.

Calibration is also a well-known technique to deal with auxiliary information in estimation. Ranalli *et al.* (2015) proposed different calibration estimators for the dual frame case, which can be easily extended to the multiple frame context. A calibration estimator in the case of more than two sampling frames can be defined as

$$\hat{P}_{CALi} = \frac{1}{N}\sum_{k\in s} z_{ki} d_k^{CAL} \tag{A4.8}$$

where d_k^{CAL} are such that they minimize $\sum_{k\in s} G\left(d_k^{CAL}, d_k^M\right)$, where $G(\cdot,\cdot)$ is a particular distance function, subject to

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

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

where $\delta_k(A_q)$ is the indicator variable that takes value 1 if unit k is in frame q and zero otherwise, and $\mathbf{t}_{xq}$ are the population totals of $\boldsymbol{x}_q$.

Recently, Elkasabi *et al.* (2015) proposed a joint calibration estimator for the dual frame case that can be easily extended to the case of three or more frames. The estimator is in the form

$$\hat{P}_{JCEi} = \frac{1}{N}\sum_{k\in s} z_{ki} d_k^{JCE} \tag{A4.9}$$

with $d_k^{JCE} = d_k(1 + \lambda'\mathbf{x}_k)$ and

$$\lambda' = \left(\sum_{k\in U}\mathbf{x}_{qk} - \sum_{k\in s} d_k\mathbf{x}_k\right)' \left(\sum_{k\in s} d_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}$$

As for $\hat{P}_{PELi}$, the same set of auxiliary variables $\boldsymbol{x} = (x_1, x_2, ..., x_p)$ is assumed to be known in all frames.

A4.3 Proposed estimators for responses with ordinal outcomes

Estimators reviewed in the previous section were originally formulated for estimating parameters (usually a total or a mean) of a continuous variable. They can be used also for estimating proportions of an ordinal variable although final estimates may likely be inconsistent, in the sense that they can take value outside the interval $[0;1]$ and they may not add up to 1. Moreover, they are not taking into account the extra information we have from the order among categories. In this case, an approach based on an ordinal logistic model (OLM) seems to be more appropriate. Within OLMs, the most widely used one 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 review on ordinal logistic models.

Under this scenario, we can exploit superpopulation models for inference from sample surveys. A superpopulation model is a way of formalizing the relationship between a target variable and auxiliary data. Superpopulation models have been used in sociological and electoral studies Cassel *et al.*(1997) used the superpopulation approach to estimate the average customer satisfaction, Pavia and Larraz (2012) used superpopulation models in electoral polls,...) Traditionally, linear regression models have been used to incorporate auxiliary information. As it is well known in sociological literature (Winship and Mare, 1984), for qualitative variables a linear model is unrealistic.

Considering the most general case, where auxiliary information differs by frame, we consider a different superpopulation ordinal logistic model in each frame. So, in frame q, the logit transformation of the cumulative probabilities can be written as follows

$$\text{logit}log(P(y_k \leq i)) = \log \frac{P(y_k \leq i)}{P(y_k > i)} = \alpha_i^q + \boldsymbol{\beta}_i^q \mathbf{x}_{qk}, \quad i = 1, ..., m-1, \quad q = 1, \ldots, Q, \tag{A4.10}$$

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

$$P(y_k \leq i) = \frac{\exp(\alpha_i^q + \boldsymbol{\beta}_i^q \mathbf{x}_{qk})}{1 + \exp(\alpha_i^q + \boldsymbol{\beta}_i^q \mathbf{x}_{qk})}, \quad i = 1, ..., m-1, \quad q = 1, \ldots, Q. \tag{A4.11}$$

We assume that, in frame q, the finite population under study $\mathbf{y} = (y_1, ..., y_N)'$ is the determination of the superpopulation random variable vector $\mathbf{Y} = (Y_1, ..., Y_N)'$, that can be described by the

superpopulation model, ξ_q, s.t.

$$\mu_i^q(\mathbf{x}_{qk}) = P(Y_k = i|\mathbf{x}_{qk}) = E_{\xi_q}(Z_{ki}|\mathbf{x}_{qk}) = \begin{cases} \frac{\exp(\alpha_i^q+\boldsymbol{\beta}_i^q\mathbf{x}_{qk})}{1+\exp(\alpha_i^q+\boldsymbol{\beta}_i^q\mathbf{x}_{qk})}, & i=1 \\ \frac{\exp(\alpha_i^q+\boldsymbol{\beta}_i^q\mathbf{x}_{qk})}{1+\exp(\alpha_i^q+\boldsymbol{\beta}_i^q\mathbf{x}_{qk})} - \frac{\exp(\alpha_{i-1}^q+\boldsymbol{\beta}_{i-1}^q\mathbf{x}_{qk})}{1+\exp(\alpha_{i-1}^q+\boldsymbol{\beta}_{i-1}^q\mathbf{x}_{qk})}, & i=2,...,m \end{cases}. \tag{A4.12}$$

Here E_{ξ_q} denotes the expected value with respect to the model in frame q and we assume that Y_k are conditionally independent given $\mathbf{x}_{qk}$. An important property that is usually assumed to be accomplished is the proportional odds property. According to this property, effects of the predictors are the same across categories. This implies that $\boldsymbol{\beta}_i^q = \boldsymbol{\beta}^q$, i.e. parameters associated to independent variables are fixed and independent of the category considered. Then, the superpopulation model can be rewritten as

$$\mu_i^q(\mathbf{x}_{qk}) = P(Y_k = i|\mathbf{x}_{qk}) = E_{\xi_q}(Z_{ki}|\mathbf{x}_{qk}) = \begin{cases} \frac{\exp(\alpha_i^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})}{1+\exp(\alpha_i^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})}, & i=1 \\ \frac{\exp(\alpha_i^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})}{1+\exp(\alpha_i^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})} - \frac{\exp(\alpha_{i-1}^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})}{1+\exp(\alpha_{i-1}^q+\boldsymbol{\beta}^q\mathbf{x}_{qk})}, & i=2,...,m \end{cases} \tag{A4.13}$$

Usually, population parameters α_i^q and $\boldsymbol{\beta}^q$ involved in the model ξ_q are unknown and should be estimated using sample information. Different procedures, as weighted least squares (Goldberger, 1964) or maximum likelihood, 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

$$\ell(\boldsymbol{\theta}^q) = \sum_{i=1,\ldots,m} \sum_{k\in s_q} d_k z_{ki} \log \mu_i^q(\mathbf{x}_{qk}, \boldsymbol{\theta}^q), \tag{A4.14}$$

and we denote it by $\hat{\boldsymbol{\theta}}^q = (\hat{\alpha}_1^q, \ldots, \hat{\alpha}_m^q, \hat{\boldsymbol{\beta}}^q)$. Under certain conditions the π-weighted log-likelihood estimator is consistent for $\boldsymbol{\theta}^q$ (Nordberg, 1989). Using these maximum likelihood estimates, we can define an estimator for probabilities for each category as follows:

$$p_{ki}^q = \hat{\mu}_i^q(\mathbf{x}_{qk}) = \begin{cases} \frac{\exp(\hat{\alpha}_i^q+\hat{\boldsymbol{\beta}}^q\mathbf{x}_{qk})}{1+\exp(\hat{\alpha}_i^q+\hat{\boldsymbol{\beta}}^q\mathbf{x}_{qk})}, & i=1 \\ \frac{\exp(\hat{\alpha}_i^q+\hat{\boldsymbol{\beta}}^q\mathbf{x}_{qk})}{1+\exp(\hat{\alpha}_i^q+\hat{\boldsymbol{\beta}}^q\mathbf{x}_{qk})} - \frac{\exp(\hat{\alpha}_{i-1}^q+\hat{\boldsymbol{\beta}}^q\mathbf{x}_{qk})}{1+\exp(\hat{\alpha}_{i-1}^q+\hat{\boldsymbol{\beta}}^q\mathbf{x}_{qk})}, & i=2,...,m \end{cases}. \tag{A4.15}$$

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

$$\hat{P}_{MA1i} = \frac{1}{N}\left(\sum_{q=1}^{Q}\sum_{k\in U}\frac{p_{ki}^{q}}{m_k} - \sum_{k\in s} p_{ki}^{q} d_k^M + \sum_{k\in s} z_{ki} d_k^M\right), \quad i = 1, ..., m \tag{A4.16}$$

$$\hat{P}_{MA2i} = \frac{1}{N}\left(\sum_{q=1}^{Q}\sum_{k\in U}\frac{p_{ki}^{q}}{m_k} - \frac{N}{\hat{M}}\sum_{k\in s} p_{ki}^{q} d_k + \sum_{k\in s} z_{ki} d_k^M\right), \quad i = 1, ..., m \tag{A4.17}$$

with $\hat{M} = \sum_{k\in s} d_k$. To formulate both estimators we have adapted the approach used by Lehtonen and Veijanen (1998a) 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 in estimator $\hat{P}_{MA1i}$ are weighted by multiplicities m_k to avoid overestimation issues. For this same reason, weights d_k^M are used in the sample sums. Such weighing is intended to make the estimator consistent in the sense that its categories add up to 1. Estimator $\hat{P}_{MA2i}$ is very similar to $\hat{P}_{MA1i}$, with the only difference of using original design weights d_k in one of the sample sums. Due to this, and to ensure the consistency of the estimator, adjustment factor $N/\hat{M}$ is used.

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.

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

$$\hat{P}_{MC1i} = \frac{1}{N}\sum_{k\in s}\frac{w_k^{\circ}}{m_k} z_{ki}, \quad i = 1, ..., m, \tag{A4.18}$$

where weights w_k° are chosen so that they minimize $\sum_{k\in s} G\left(w_k^{\circ}, d_k\right)$, subject to

$$\sum_{k\in s}\frac{w_k^{\circ}}{m_k}\delta_k(A_q) = N_q, \quad q = 1, ..., Q$$

$$\sum_{k\in s}\frac{w_k^{\circ}}{m_k} p_{ki}^{q}\delta_k(A_q) = \sum_{k\in U} p_{ki}^{q}\delta_k(A_q), \quad q = 1, ..., Q, \quad i = 1, ..., m.$$

In the first group of Q calibration constraints, regarding frame sizes, multiplicities m_k are used to properly weight indicator variables $\delta_k(A_q)$ and so, to cancel any overestimation problem. The same reasoning may be applied to the second group of constraints, where the auxiliary variables are also weighted by m_k.

Alternatively to (A4.14), model parameters for the q-th frame can be estimated maximizing the following loglikelihood function

$$\ell(\boldsymbol{\theta}^q) = \sum_{i=1,\ldots,m} \sum_{k \in s_q} d_k^M z_{ki} \log \mu_i^q(\mathbf{x}_{qk}, \boldsymbol{\theta}^q), \tag{A4.19}$$

yielding to the probability estimates

$$p_{ki}^{\star q} = \hat{\mu}_i^q(\mathbf{x}_{qk}) = \begin{cases} \frac{\exp(\hat{\alpha}_i^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})}{1+\exp(\hat{\alpha}_i^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})}, & i = 1 \\ \frac{\exp(\hat{\alpha}_i^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})}{1+\exp(\hat{\alpha}_i^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})} - \frac{\exp(\hat{\alpha}_{i-1}^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})}{1+\exp(\hat{\alpha}_{i-1}^{\star q} + \hat{\boldsymbol{\beta}}^{\star q} \mathbf{x}_{qk})}, & i = 2, ..., m \end{cases}. \tag{A4.20}$$

The following calibration estimator can be defined

$$\hat{P}_{MC2i} = \frac{1}{N} \sum_{k \in s} w_k^\star z_{ki}, \quad i = 1, ..., m \tag{A4.21}$$

where, in this case, the weights $w_k^\star$ are such that they minimize $\sum_{k \in s} G\left(w_k^\star, d_k^M\right)$ subject to

$$\sum_{k \in s} w_k^\star \delta_k(A_q) = N_q, \quad q = 1, \ldots, Q$$

$$\sum_{k \in s} w_k^\star p_{ki}^{\star q} \delta_k(A_q) = \sum_{k \in U} p_{ki}^{\star q} \delta_k(A_q), \quad q = 1, \ldots, Q, \quad i = 1, ..., m.$$

Unlike those in $\hat{P}_{MC1i}$, constraints for this calibration estimator do not involve multiplicities. Over-estimation issues are eliminated, then, by considering d_k^M (which are already weighted by m_k) as the starting weights for the calibration. Therefore, resulting weights $w_k^\star$ should be near to those starting weights so they already take into account the multiplicity while still fulfilling the calibration constraints.

A4.4 Properties of the proposed estimators

In this section we describe the main properties of the proposed estimators. We adapt the asymptotic framework of Isaki and Fuller (1982) to a multiple frame context, in which 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 all domains D_K being non-empty, $K \subseteq \{1, ..., q, .., Q\}$. Additionally, $n_{q_N}/n_N \to c_q \in (0,1), q = 1, ..., Q$, 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 $\hat{P}_{MC2}$ and then move to the other estimators, because these can be dealt with using slight modifications of this more general case. Let $\mu_i^q(\mathbf{x}_{qk}, \boldsymbol{\theta}^q) = \frac{\exp(\alpha_i^q + \boldsymbol{\beta}^q \mathbf{x}_{qk})}{1+\exp(\alpha_i^q + \boldsymbol{\beta}^q \mathbf{x}_{qk})} - \frac{\exp(\alpha_{i-1}^q + \boldsymbol{\beta}^q \mathbf{x}_{qk})}{1+\exp(\alpha_{i-1}^q + \boldsymbol{\beta}^q \mathbf{x}_{qk})}$, for $i = 1, ..., m$ and $q = 1, ..., Q$. In order to prove our results, we make a set of technical assumptions reported in Appendix A4.7.1.

Theorem 2. *Under assumptions A6–A11, estimator $\widehat{P}_{MC2i}$ is design $\sqrt{n_N}$-consistent for P_i in the sense that*

$$\widehat{P}_{MC2i} - P_i = O_p(n_N^{-1/2}),$$

and has the following asymptotic distribution

$$\frac{\widehat{P}_{MC2i} - P_i}{\sqrt{V_\infty(\widehat{P}_{MC2i})}} \xrightarrow{\mathcal{L}} N(0,1)$$

where

$$V_\infty(\widehat{P}_{MC2i}) = \frac{1}{N^2} \sum_{q=1}^{Q} \left(\sum_{k \in U_q} \sum_{l \in U_q} \Delta_{kl} (d_k^M e_{ki}^q)(d_l^M e_{li}^q) \right) \tag{A4.22}$$

with $\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$ and $e_{ki}^q = z_{ki} - \mu_i^q(\mathbf{x}_{qk}, \boldsymbol{\theta}_U^q) \boldsymbol{B}_{iU}^q$, and $\boldsymbol{B}_{iU}^q = \sum_{k \in U_q} (\mu_i^q(\mathbf{x}_{qk}, \boldsymbol{\theta}_U^q)^2)^{-1} (\sum_{k \in U_q} \mu_i^q(\mathbf{x}_{qk} \boldsymbol{\theta}_U^q) z_{ki})$. In addition, let $\hat{e}_{ki}^q = z_{ki} - p_{ki}^q \hat{\boldsymbol{B}}_i^q$, being $\hat{\boldsymbol{B}}_i^q = (\sum_{k \in s_q} d_k^M p_{ki}^{q2})^{-1} (\sum_{k \in s} d_k^M p_{ki}^q z_{ki})$. Then, $V_\infty(\widehat{P}_{MC2i})$ can be consistently estimated by

$$v(\widehat{P}_{MC2i}) = \frac{1}{N^2} \sum_{q=1}^{Q} \left(\sum_{k \in s_q} \sum_{l \in s_q} (\frac{\Delta_{kl}}{\pi_{kl}}) (d_k^M \hat{e}_{ki}^q)(d_l^M \hat{e}_{li}^q) \right)$$

Proof. See Appendix A4.7.2

Estimator $\hat{P}_{MC1}$ is similar to $\hat{P}_{MC2}$. The only differences are (i) $\hat{P}_{MC1}$ uses original design weights d_k as starting weights for the calibration, correcting with multiplicities m_k where necessary to avoid overestimating issues and (ii) probability estimates p_{ki}^q are used as auxiliary information instead of $p_{ki}^{\star q}$.

On the other hand, $\hat{P}_{MA1}$ and $\hat{P}_{MA2}$ can be seen as particular cases of $\widehat{P}_{MC2i}$ in which $\boldsymbol{B}_{iU}^{q}$ is a scalar equal to 1 for all i and all q and, again, estimates p_{ki}^{q} are used as auxiliary information. In both estimators population sum of probabilities is weighted by m_k to correct for the multiplicity. The main difference between estimator $\hat{P}_{MA2}$ and estimator $\hat{P}_{MA1}$ is that in $\hat{P}_{MA2}$ the term $\sum_{q=1}^{Q}\sum_{k\in U_q}\mu_i^q(\mathbf{x}_{qk})$ is estimated by $\frac{N}{\hat{M}}\sum_{k\in s}p_{ki}^{q}d_k$ instead of by $\sum_{k\in s}p_{ki}^{q}d_k^M$ as in the latter. Despite of these particularities, a similar procedure to the one used with $\hat{P}_{MC2}$ can be considered to prove the results.

A4.5 Monte Carlo Simulation Experiments

We now compare empirically the performance of the proposed estimators with respect to alternative estimators via Monte Carlo experiments, which have been carried out by using the freeware statistical program R.

We have considered a three frame setting, say frames A, B and C, 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 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 ratios of the levels of 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 these domains. As a result, three overlapping frames of sizes $N_A = 5500, N_B = 6000$ and $N_C = 5000$ were obtained. Three samples of sizes $n_A = 360, n_B = 464$ and $n_C = 728$ were independently drawn, one from each frame, considering Midzuno sampling designs in frames A and C and a simple random sampling design in frame B. Sample from frame A was drawn with probabilities proportional to a normally distributed variable with mean 1000 and standard deviation 250. On the other hand, sample from frame C was drawn considering inclusion probabilities proportional to another normally distributed variables with mean 5000 and standard deviation 500. In this scenario, the two ordinal model-assisted estimators (PMA1 and PMA2) and the two ordinal model-calibrated estimators (PMC1 and PMC2) 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, PMA1, PMA2, PMC1 and PMC2) we have considered different sets of variables: x_1 in frame A, x_2 in frame B and both x_1 and x_2 in frame C.

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 1000 simulation runs. We have used $RMSE\%$ to calculate percent relative efficiency gain with respect to multiplicity estimator (results are presented in Table A4.1).

Table A4.1: % 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 (A4.5)	*-0.08*	*0.19*	*0.04*	*-0.10*	*0.04*	*0.19*	*0.10*
	100.00	100.00	100.00	100.00	100.00	100.00	100.00
KA (A4.3)	*-0.14*	*0.16*	*0.05*	*-0.09*	*0.05*	*0.16*	*0.11*
	107.00	104.11	104.47	104.26	104.11	107.00	104.96
CM (A4.6)	*-0.14*	*0.16*	*0.07*	*-0.10*	*0.07*	*0.16*	*0.12*
	106.71	103.84	104.16	103.92	103.84	106.71	104.65
CAL (A4.8)	*-0.43*	*0.24*	*0.21*	*-0.17*	*0.17*	*0.43*	*0.26*
	134.47	113.68	99.41	173.16	99.41	173.16	130.18
PMA1 (A4.16)	*0.71*	*-0.12*	*-0.31*	*0.11*	*0.11*	*0.71*	*0.31*
	190.77	132.25	119.69	216.48	119.69	216.48	164.79
PMA2 (A4.17)	*0.09*	*0.16*	*-0.01*	*-0.10*	*0.01*	*0.16*	*0.09*
	166.77	122.88	114.08	179.71	114.08	179.71	145.86
PMC1 (A4.18)	*-0.08*	*0.13*	*0.03*	*-0.07*	*0.03*	*0.13*	*0.08*
	183.79	129.51	121.54	192.99	121.54	192.99	156.95
PMC2 (A4.21)	*-0.10*	*0.14*	*0.02*	*-0.06*	*0.02*	*0.14*	*0.08*
	184.70	129.50	121.75	195.23	121.75	195.23	157.79

From results of table A4.1 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. All the proposed ordinal estimators work better than the classical

calibration estimator, which assume an underlying linear model. Whatever the proposed estimators, 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. Within the group of proposed estimators, PMA1 is the estimator which shows the largest efficiency gain. On the other hand, no significative differences can be detected between the two calibration estimators proposed.

To determine the effect of varying association between response and auxiliary variables, we are going to consider new scenarios with different correlation levels between y and x_1 and x_2. In the first scenario created, correlation between y and x_1 has been decreased with respect to the initial situation to 0.65. On the other hand, correlation between y and x_2 has been set to 0.5. In the second scenario, correlation levels between y and x_1 and between y and x_2 are set to 0.4 and 0.7, respectively. We have run 1000 repetitions 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 A4.2.

Table A4.2: % 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
				$\rho_{YX_1} = 0.65$, $\rho_{YX_2} = 0.5$.			
PMA1	125.15	110.56	104.16	133.48	104.16	133.48	118.33
PMA2	119.00	106.49	102.14	123.58	102.14	123.58	112.80
PMC1	126.53	107.40	103.46	130.77	103.46	130.77	117.04
PMC2	125.96	107.18	103.21	130.84	103.21	130.84	116.79
				$\rho_{YX_1} = 0.4$, $\rho_{YX_2} = 0.7$.			
PMA1	122.54	110.27	106.44	133.53	106.44	133.53	118.19
PMA2	116.57	105.99	103.36	124.20	103.36	124.20	112.53
PMC1	124.35	107.16	104.99	131.43	104.99	131.43	116.98
PMC2	123.59	106.97	104.80	131.60	104.80	131.60	116.74

We observe that 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.

A4.5.1 Application to real data

In addition to simulation studies, we have utilized a set of real data to check the performance of the proposed estimators. 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) and another one of cell phones (frame B). Finally, $n = 1853$ 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 region of Andalusia and then a stratified sample of size $n_A = 1468$ was drawn. In cell phone frame a simple random sample of size $n_B = 385$ was selected by using a random digit dialing (RDD) method.

We have considered two different response variables related with attitudes regarding immigration. The first one is "The place you prefer for living is a place with..." where possible options are "...few immigrants", "...some immigrants" or "...many immigrants". The second main variable is the response to question "Do you consider that immigrants have nothing, little, quite a few or much in common with you?" As auxiliary information we use the sex and the age (categorized by considering 4 age classes) of interviewed people in each frame. Population data for auxiliary variables is available from table A4.3.

Together with the proposed estimators, we have calculated some additional estimators for comparison purposes as multiplicity (M), Kalton-Anderson (KA), composite multiplicity (CM), calibration (CAL)) and joint calibration (JCE) estimators. For CAL and JCE we have used also the sex and the age of the individuals as auxiliary variables to get comparable results. Note these estimators are the alternatives available given the same amount of auxiliary information but both estimators work well when the relationship between main and auxiliary variable is strongly linear.

Table A4.4 shows point estimation for compared estimators for the two main variables. We have used the jackknife procedure described in Lohr and Rao (2000) to estimate variance of estimators and then a 95 % confidence interval has been computed. Results of lower bound, upper bound and lengths of intervals are also includes in the table.

In both cases, average length of confidence intervals of all proposed estimators is smaller than average

Table A4.3: Population data for variables `sex` and `age`

	Both	Landline	Cell
	Men		
18 - 29	428750	0	188172
30 - 44	724435	4259	298416
45 - 59	603338	59385	135981
> 60	396626	206410	94729
	Women		
18 - 29	480151	0	115472
30 - 44	658984	17673	289106
45 - 59	601478	39362	141553
> 60	445897	316172	104567

lengths of confidence intervals of classical estimators.

A4.6 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 proprieties 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 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 ordinal model-assisted estimators provide all a sensible reduction on the length on the confidence intervals for the estimated proportions compared to all other estimators.

A4.7 Appendix - Assumptions and proof of Theorem 4.1.

A4.7.1 Assumptions

A6. *Let* $\boldsymbol{\theta}_U^q$ *be the census level parameter estimate obtained by maximizing the likelihood*

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

Assume that $\boldsymbol{\theta}^q = \lim_{N\to\infty} \boldsymbol{\theta}_U^q$ *exists and that* $\hat{\boldsymbol{\theta}}^q = \boldsymbol{\theta}_U^q + O_p(n_N^{-1/2})$, $q = 1, \ldots, Q$.

A7. *For each* $\mathbf{x}_{qk}$, $\partial\mu(\mathbf{x}_{qk}, \boldsymbol{t})/\partial \boldsymbol{t}$ *is continuous in* $\boldsymbol{t}$ *and* $|\partial\mu(\mathbf{x}_{qk}, \boldsymbol{t})/\partial \boldsymbol{t}| \leq f_1(\mathbf{x}_{qk}, \boldsymbol{\theta}^q)$ *for* $\boldsymbol{t}$ *in a neighborhood of* $\boldsymbol{\theta}^q$ *and* $f_1(\mathbf{x}_{qk}, \boldsymbol{\theta}^q) = O(1)$, *for* $i = 1, \ldots, m; q = 1, \ldots, Q$.

A8. *For each* $\mathbf{x}_{qk}$, $\partial^2\mu(\mathbf{x}_{qk}, \boldsymbol{t})/\partial t_j t_{j'}$ *is continuous in* $\boldsymbol{t}$ *and* $\max_{j,j'} |\partial^2\mu(\mathbf{x}_{qk}, \boldsymbol{t})/\partial t_j t_{j'}| \leq f_2(\mathbf{x}_{qk}, \boldsymbol{\theta}^q)$ *for* $\boldsymbol{t}$ *in a neighborhood of* $\boldsymbol{\theta}^q$ *and* $f_2(\mathbf{x}_{qk}, \boldsymbol{\theta}^q) = O(1)$, *for* $i = 1, \ldots, m; q = 1, \ldots, Q$.

A9. *The auxiliary variables* $\boldsymbol{x}$ *have bounded fourth moments.*

A10. *For any study variable* ξ *with bounded fourth moment, the sampling designs are such that for the normalized multiplicity estimators of* $\bar{\xi} = N^{-1} \sum_{k\in U} \xi_k$ *a central limit theorem holds, i.e.*

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

where $\hat{\bar{\xi}}_M = N^{-1} \sum_{k\in s} d_k^M \xi_k$.

A11. *Let* $\boldsymbol{B}_{iU}^q = \sum_{k\in U_q} (\mu_i^q(\mathbf{x}_{qk}, \boldsymbol{\theta}_U^q)^2)^{-1} (\sum_{k\in U_q} \mu_i^q(\mathbf{x}_{qk}, \boldsymbol{\theta}_U^q) z_{ki})$. *Assume that* $\boldsymbol{B}_i^q = \lim_{N\to\infty} \boldsymbol{B}_{iU}^q$ *exists, and the sampling designs are such that* $\boldsymbol{B}_{iU}^q$ *can be consistently estimated by* $\hat{\boldsymbol{B}}_i^q$ *for* $i = 1, ..., m; q = 1, ..., Q$.

A4.7.2 Proof of Theorem 4.1.

Estimator $\hat{P}_{MC2i}$ can be rewritten in the form

$$\hat{P}_{MC2i} = \frac{1}{N}\sum_{k\in s} d_k^M z_{ki} + \frac{1}{N}\sum_{q=1}^{Q}\left(\sum_{k\in U}\boldsymbol{p}_{ki}^{\star q} - \sum_{k\in s} d_k^M \boldsymbol{p}_{ki}^{\star q}\right)\boldsymbol{B}_{iU}^q + \frac{1}{N}\sum_{q=1}^{Q}\left(\sum_{k\in U}\boldsymbol{p}_{ki}^{\star q} - \frac{1}{N}\sum_{k\in s_q} d_k^M \boldsymbol{p}_{ki}^{\star q}\right)(\hat{\boldsymbol{B}}_i^q - \boldsymbol{B}_{iU}^q)$$

with $\boldsymbol{p}_{ki}^{\star q} = (\delta_k(A_q), p_{ki}^{\star q}\delta_k(A_q))$.

Now, using the same approach developed in Montanari and Ranalli (2005), it is easy to show that by assumption A11, $\hat{\boldsymbol{B}}_i^q - \boldsymbol{B}_{iU}^q = o(1)$; and by assumptions A6–A7 and A9–A10

$$\frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in U} p_{ki}^{\star q}\delta_k(A_q) - \frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in s} d_k^M p_{ki}^{\star q}\delta_k(A_q) = O_p(n^{-1/2}),$$

using a first order Taylor expansion of $\mu(\boldsymbol{x}_{qx}, \hat{\boldsymbol{\theta}}^q)$ at $\hat{\boldsymbol{\theta}}^q = \boldsymbol{\theta}_U^q$. Using A6–A10 and a second order Taylor expansion of $\mu(\boldsymbol{x}_{qx}, \hat{\boldsymbol{\theta}}^q)$ at $\hat{\boldsymbol{\theta}}^q = \boldsymbol{\theta}_U^q$

$$\frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in U} p_{ki}^{\star q}\delta_k(A_q) - \frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in s} d_k^M p_{ki}^{\star q}\delta_k(A_q) = \frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in U} \mu_{ki}^q\delta_k(A_q) - \frac{1}{N}\sum_{q=1}^{Q}\sum_{k\in s} d_k^M \mu_{ki}^q\delta_k(A_q) = O_p(n^{-1})$$

Thus,

$$\hat{P}_{MC2i} = \frac{1}{N}\sum_{k\in s} d_k^M z_{ki} + \frac{1}{N}\sum_{q=1}^{Q}\left(\sum_{k\in U}\boldsymbol{\mu}_{ki}^{\star q} - \sum_{k\in s} d_k^M \boldsymbol{\mu}_{ki}^{\star q}\right)\boldsymbol{B}_{iU}^q + o_p(n^{-1}),$$

where $\boldsymbol{\mu}_{ki}^{\star q}$ is like $\boldsymbol{p}_{ki}^{\star q}$ but with $p_{ki}^{\star q}$ replaced by its population counterpart $\mu_{ki}^{\star q}$. Consequently

$$E_p(\hat{P}_{MC2i}) \to E_p(\frac{1}{N}\sum_{k\in s} d_k^M z_{ki}) = P_i$$

and

$$V_p(\hat{P}_{MC2i}) \to V_p(\frac{1}{N}\sum_{k\in s} d_k^M (z_{ki} - \mu_i^q(\boldsymbol{x}_{qk}, \boldsymbol{\theta}_U^q))\boldsymbol{B}_{iU}^q).$$

Under assumption A10, estimator $\frac{1}{N}\sum_{k\in s} d_k^M (z_{ki} - \mu_i^q(\boldsymbol{x}_{qk}, \boldsymbol{\theta}_U^q))\boldsymbol{B}_{iU}^q$ is asymptotic normal distributed,

so we can conclude that estimator $\hat{P}_{MC1i}$ is also asymptotic normal distributed.

Table A4.4: Point and 95% confidence level estimation of percentages using Jackknife variance estimation. Auxiliary variables: Sex and Age.

The place you prefer for living is a place with...

Estimator	PROP	LB	UB	LEN
		...few immigrants		
M	45.26	41.93	48.58	6.65
KA	44.70	41.60	47.78	6.17
CM	44.92	41.84	48.00	6.17
CAL	44.43	41.47	47.39	5.92
JCE	44.97	42.03	47.90	5.87
PMA1	45.29	42.24	48.32	6.07
PMA2	45.60	42.33	48.86	6.53
PMC1	44.55	41.67	47.42	5.75
PMC2	44.67	41.73	47.60	5.87
		...some immigrants		
M	48.36	45.03	51.67	6.64
KA	49.18	46.07	52.27	6.20
CM	48.65	45.56	51.72	6.16
CAL	49.46	46.45	52.48	6.03
JCE	48.80	45.86	51.72	5.86
PMA1	48.51	45.47	51.54	6.07
PMA2	48.13	44.86	51.39	6.53
PMC1	49.50	46.55	52.45	5.90
PMC2	49.30	46.22	52.37	6.15
		...many immigrants		
M	6.39	4.68	6.38	3.41
KA	6.13	4.62	7.62	3.00
CM	6.43	4.83	6.43	3.20
CAL	6.09	4.68	7.51	2.83
JCE	6.23	4.78	7.67	2.89
PMA1	6.20	4.67	7.72	3.06
PMA2	6.27	4.60	7.93	3.32
PMC1	5.95	4.56	7.33	2.76
PMC2	6.03	4.58	7.47	2.89

Do you consider that immigrants have nothing, little, something, quite a few or much in common with you?

Estimator	PROP	LB	UB	LEN
		Nothing		
M	12.13	9.80	14.45	4.65
KA	11.09	9.08	13.08	4.01
CM	12.15	9.96	14.34	4.38
CAL	11.37	9.41	13.33	3.92
JCE	11.34	9.47	13.20	3.73
PMA1	12.18	10.04	14.31	4.27
PMA2	12.50	10.17	14.82	4.65
PMC1	11.41	9.54	13.26	3.71
PMC2	11.63	9.64	13.60	3.96
		Little		
M	27.87	24.94	30.80	5.87
KA	27.92	25.20	30.63	5.43
CM	28.02	25.29	30.74	5.45
CAL	28.59	25.92	31.25	5.33
JCE	28.60	25.96	31.24	5.28
PMA1	28.73	26.03	31.43	5.40
PMA2	28.45	25.55	31.33	5.78
PMC1	28.42	25.83	31.00	5.16
PMC2	28.39	25.72	31.06	5.34
		Something		
M	10.86	8.78	12.93	4.15
KA	10.99	9.03	12.93	3.90
CM	10.65	8.74	12.56	3.82
CAL	10.83	9.00	12.65	3.65
JCE	10.47	8.68	12.26	3.58
PMA1	10.75	8.86	12.62	3.76
PMA2	10.81	8.78	12.83	4.05
PMC1	10.80	9.01	12.58	3.57
PMC2	10.79	8.99	12.59	3.60
		Quite a few		
M	29.30	26.29	32.29	5.99
KA	29.82	27.00	32.63	5.62
CM	29.33	26.54	32.10	5.56
CAL	29.09	26.42	31.75	5.33
JCE	28.94	26.34	31.54	5.20
PMA1	28.90	26.22	31.58	5.36
PMA2	28.88	25.98	31.77	5.78
PMC1	29.14	26.54	31.73	5.18
PMC2	29.04	26.43	31.65	5.22
		Much		
M	19.84	17.22	22.45	5.23
KA	20.19	17.72	22.65	4.93
CM	19.85	17.43	22.25	4.83
CAL	20.11	17.76	22.47	4.71
JCE	20.64	18.22	23.05	4.82
PMA1	19.43	17.08	21.78	4.70
PMA2	19.36	16.82	21.90	5.08
PMC1	20.23	17.90	22.56	4.67
PMC2	20.14	17.78	22.50	4.72

The place you prefer for living is a place with...				
Estimator	REDUCTION			
	...few immigrants	*...some immigrants*	*...many immigrants*	MEAN
PMA1	8.65	8.66	10.40	9.24
PMA2	1.86	1.76	2.67	2.10
PMC1	13.50	11.26	19.04	14.60
PMC2	11.79	7.40	15.33	11.51

Table A4.5: Relative length reduction in % of the 95% confidence intervals of the proposed estimators with respect to the multiplicity estimator.

Do you consider that immigrants have nothing, little, something, quite a few or much in common with you?						
Estimator	REDUCTION					
	Nothing	*Little*	*Something*	*Quite a few*	*Much*	MEAN
PMA1	8.15	7.94	9.43	10.52	10.23	9.25
PMA2	-0.05	1.44	2.45	3.50	2.96	2.06
PMC1	20.09	11.98	13.98	13.52	10.79	14.07
PMC2	14.78	9.03	13.33	12.90	9.72	11.95

Table A4.6: Relative length reduction in % of the 95% confidence intervals of the proposed estimators with respect to the multiplicity estimator.

Bibliography

[1] Abascal, E., Díaz de Rada, V., García, I. and Landaluce, I. (2012). Face to face and telephone surveys in terms of sampling representativeness: a multidimensional analysis. *Quality and Quantity*, **46**, 303–313.

[2] Agresti, A. (2007). *An introduction to categorical data analysis*. 2nd Edition. Hoboken: Wiley Series in Probability and Statistics. Wiley-Interscience

[3] Alfons, A., Holzer, J and Templ, M. (2014). *laeken: estimation of indicators on social exclusion and poverty*. URL `http://CRAN.R-project.org/package=laeken`. R package version 0.4.6.

[4] Arcos, A., Rueda, M., Ranalli, M. G. and Molina, D. (2015). *Frames2: Estimation in dual frame surveys*. URL `http://CRAN.R-project.org/package=Frames2`. R package version 0.1.1.

[5] Arcos, A., Molina, D., Rueda, M. and Ranalli, M. G. (2015). Frames2: A package for estimation in dual frame surveys. *The R Journal*, **7**(1), 52–72.

[6] 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.

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

[8] Brick, J. M., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, **70**(5), 780–793.

[9] Brick, J. M., Edwards, W. S. and Lee, S. (2007). Sampling Telephone Numbers and Adults, Interview Length, and Weighting in the California Health Interview Survey Cell Phone Pilot Study. *Public Opinion Quarterly*, **71**(5), 793–813.

[10] Busse, B. and Fuchs, M. (2012). The components of landline telephone survey coverage bias. The relative importance of no-phone and mobile-only populations. *Quality and Quantity*, **46**, 1209–1225.

[11] Cassel, C. M., Särndal, C. E., and J. H. Wretman. (1997). *Foundations of inference in survey sampling.* New York: Wiley.

[12] Chen, S. and Kim, J. K. (2014). Population empirical likelihood for nonparametric inference in survey sampling. *Statistica Sinica*, **24**, 335 – 355.

[13] Couper, M. P. (2000). Web surveys: a review of issues and approaches. *Public Opinion Quarterly*, **64**, 464 – 494.

[14] Deville, J. C. (1993). Estimation de la variance pour les enquêtes en deux phases. *Manuscript*, INSEE, Paris.

[15] Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling *Journal of the American Statistical Association*, **87**(418), 376 – 382.

[16] Díaz de Rada, V. (2011). Face-to-face versus telephone surveys on political attitudes: a comparative analysis. *Quality and Quantity*, **45**(4), 817–827.

[17] Elksabai, M. A., Heeringa, S. G. and Lepkowski, J. M. (2015). Joint calibration estimator for dual frame surveys *Statistics in Transition*, **16**(1), 7–36.

[18] 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 Sections*, 245–249.

[19] Fuller, W. A, Kennedy, W., Schell, D., Sullivan, G. and Park, H. J. (1989). *PCCARP*. Iowa State University Statistical Laboratory, 1989.

[20] Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Internat. Statist. Rev.*, **54**(2), 127–138.

[21] Goldberger, A. S. (1964). *Econometric theory.* New York: Wiley.

[22] González-Villalobos, A. and Wallace M. A. (1996). Multiple frame agriculture surveys, vol. 1 and 2. Food and Agriculture Organization of the United Nations, Rome.

[23] Gutierrez Rojas, H. A. (2014). *TeachingSampling: Selection of samples and parameter estimation in finite population.* URL `http://CRAN.R-project.org/package=TeachingSampling`. R package version 3.2.1.

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

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

[26] Iachan, R. and Dennis, M. L. (1993). A multiple frame approach to sampling the homeless and transient population. *Journal of Official Statistics*, **9**(4), 747 – 764.

[27] IBM Corporation. (2013). *IBM SPSS Statistics for Windows, Version 22.0.* Armonk, NY. URL `http://www.ibm.com/`.

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

[29] Jackson, A. C., Pennay, D., Dowling, N. A., Coles-Janess, B. and Christensen, D. R. (2014). Improving gambling survey research using dual-frame sampling of landline and mobile phone numbers. *Journal of Gambling Studies*, **30**(2), 291 – 307.

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

[31] Kennedy, C. (2007). Evaluating the effects of screening for telephone service in dual frame RDD surveys. *Public Opinion Quarterly*, **71**, 750–771.

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

[33] Lehtonen, R. and Veijanen. A. (1998b). Logistic generalized regression estimators. *Survey Methodology* 24: 51-55.

[34] Lepkowski, J. M., Tucker, C., Brick, J. M., de Leeuw, E., Japec, L., Lavrakas, P. J., Link, M. W. and Sangster, R. L. eds. (2007). *Advances in telephone survey methodology*. New York: J.W. Wiley and Sons, Inc.

[35] Lohr, S. L. (2007). Recent developments in multiple frame surveys. In *Joint Statistical Meeting of the American Statistical Association*, 3257–3264.

[36] Lohr, S. L. (2009a). *Sampling: design and analysis*. 2nd Edition. Boston: Brooks/Cole.

[37] Lohr, S. L. (2009b). *Multiple frame surveys*. In D. Pfeffermann and C. R. Rao (Eds). *Handbook of Statistics: Vol. 29A. Sample surveys: Design, methods and applications* (pp. 71–78). Amsterdam: North Holland.

[38] Lohr, S. (2010). Dual frame surveys: recent developments and challenges, paper presented at the *45th Scientific Meeting of the Italian Statistical Society*, Padua, Italy, June 16 - 18.

[39] Lohr, S. and Rao, J. N. K. (2000). Inference in dual frame surveys. *Journal of the American Statistical Association*, **95**(449), 271–280.

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

[41] Lohr, S. L. and Brick, J. M. (2014). Allocation for dual frame telephone surveys with nonresponse. *Journal of Survey Statistics and Methodology*, **2**(4), 388–409.

[42] Lopez, M. H., Gonzalez-Barrera, A. and Krogstad, J. M. (2014). Latino support for democrats falls, but democratic advantage remains: immigration not a deal-breaker issue for half of latino voters. Washington, D.C.: Pew Research Center, October.

[43] Lu, B., Sahr, T., Iachan, R., Denker, M., Duffy, T. and Weston, D. (2013). Design and analysis of dual-frame telephone surveys for health policy research. *World Medical & Health Policy*, **5**(3), 217–232.

[44] Lumley, T. (2014). *survey: analysis of comples survey samples*. URL `http://CRAN.R-project.org/package=survey`. R package version 3.30.

[45] Lund, R. E. (1968). Estimators in multiple frame surveys. In *Proceedings of the American Statistical Association, Social Statistics Sections*, 282–288.

[46] McMillen, R. C., Winickoff, J. P., Wilson, K., Tanski, S. and Klein, J. D. (2015). A dual-frame sampling methodology to address landline replacemente in tobacco control research. *Tobacco Control*, **24**(1), 7 – 10.

[47] Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology*, **33**, 151–158.

[48] Mecatti, F. and Singh, A. C. (2014). Estimation in multiple frame surveys: a simplified and unified review using the multiplicity approach. *Journal de la Société Franaise de Statistique*, **155**(4), 51 – 69.

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

[50] Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, **100**(472), 1429–1442.

[51] Nordberg, L. (1989). Generalized Linear Modeling of Sample Survey Data. *Journal of Official Statistics* **5**, 223-239.

[52] Opsomer, J. (2011). Innovations in survey sampling design: discussion of three contributions presented at the U.S. Census Bureau. *Survey Methodology*, **37**(2), 227–231.

[53] Pasadas, S., Trujillo, M., Sánchez, A. and Reche, J. L. (2011). La incorporación de las líneas móviles al marco muestral de las encuestas telefónicas: Pertinencia, métodos y resultados. *Metodología de Encuestas*, **13**, 33–54.

[54] Pasadas, S. and Trujillo, M. (2013). Afijación óptima basada en costes para muestras telefónicas recogidas en marcos duales. In *1st Southern European Conference on Survey Methodology (SESM) and VI Congreso de Metodología de Encuestas* Barcelona, 12th - 14th December.

[55] Pavía, J. M. and Larraz, B. (2012) Nonresponse Bias and Superpopulation Models in Electoral Polls *Reis* **137**, 121-150.

[56] Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, **20**, 355–375.

[57] Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, **43**, 353–360.

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

[59] Rao, J. N. K. and Skinner, C. J. (1996). Estimation in dual frame surveys with complex designs. In *Proceedings of the Survey Method Section, Statistical Society of Canada*, 63–68.

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

[61] Research Triangle Institute. (2013). *SUDAAN, Version 11.0.1*. Research Triangle Park, NC. URL `http://www.rti.org/sudaan/`.

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

[63] Särndal, C. E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

[64] SAS Institute Inc. (2013). *SAS Software, Version 9.4*. Cary, NC. URL `http://www.sas.com/`.

[65] Singh, A. C. and Mecatti, F. (2011). Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *Journal of Official Statistics 27*(4), 633–650.

[66] Skinner, C. J. (1991). On the efficiency of raking ration estimation for multiple frame surveys. *Journal of the American Statistical Association*, **86**, 779–784.

[67] Skinner, C. J. and Rao, J. N. K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, **91**(433), 349–356.

[68] Stata Corporation. (2015). *Stata Statistical Software, Version 14.0*. College Station, TX. URL `http://www.stata.com/`.

[69] Statistics Canada. Health Statistics Division. (2005) Canadian Community Health Survey, 2003.

[70] Systat Software Inc. (2009). *Systat, Version 13.0*. San Jose, California. URL `http://www.systat.com`.

[71] Templ, M. (2014). *CRAN Task View: Official statistics and Survey methodology*. URL `http://cran.r-project.org/web/views/OfficialStatistics.html`.

[72] Tillé, Y. and Matei, A. (2012) *sampling: Survey Sampling*. URL `http://CRAN.R-project.org/package=sampling`. R package version 2.5.

[73] Trujillo M., Domínguez, J. A. and Pasadas, S. (2005). Mobile phones and their impacts on survey data. In *European Association for Survey Research Conference* Barcelona, 18th - 22th July.

[74] Vicente, P. and Reis, E. (2009). The mobile-only population in Portugal and its impact in a dual frame telephone survey. *Survey Research Methods*, **3**(2), 105–111.

[75] Vicente, P., Reis, E. and Santos, M. (2009). Using mobile phones for survey research: a comparison with fixed phones. *International Journal of Market Research*, **51**(5), 613–633.

[76] Winship C, Mare R. D. (1984). Regression Models with Ordinal Variables. *American Sociological Review* [Internet].

[77] Wolter, K. M. (2007). *Introduction to variance estimation*. 2nd Edition. New York: Springer, Inc.

[78] Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Canandian Journal of Statistics*, **32**, 15–26.

[79] Wu, C. (2005). Algorithms and R codes for the pseudo empirical likelihood method in survey sampling. *Survey Methodology*, **31**(2), 239–243.

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

[81] Wu, C. and Sitter, R. R. (2001a). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, **96**(453), 185–193.

[82] Wu, C. and Sitter, R. R. (2001b). Variance estimation for the finite population distribution function with complete auxiliary information. *Canadian Journal of Statististics*, **29**(2), 289–307.