Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

The estimation of gross flows in complex surveys with random nonresponse

by Andrés Gutiérrez, Leonardo Trujillo and Pedro Luis do Nascimento Silva

Release date: December 19, 2014

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 12-001-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by “Key resource” > “Publications.”

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “About us” > “The agency” > “Providing services to Canadians.”

Published by authority of the Minister responsible for Statistics Canada

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Survey Methodology, December 2014
Vol. 40, No. 2, pp. 285-321
Statistics Canada, Catalogue No. 12-001-X

The estimation of gross flows in complex surveys with random nonresponse

Andrés Gutiérrez, Leonardo Trujillo and Pedro Luis do Nascimento Silva[1]

Abstract

Rotating panel surveys are used to calculate estimates of gross flows between two consecutive periods of measurement. This paper considers a general procedure for the estimation of gross flows when the rotating panel survey has been generated from a complex survey design with random nonresponse. A pseudo maximum likelihood approach is considered through a two-stage model of Markov chains for the allocation of individuals among the categories in the survey and for modeling for nonresponse.

Key Words: Design-based inference; Rotating panel surveys; Gross flows; Markov chains.

1 Introduction

Survey techniques are commonly used in order to estimate some parameters of interest in a finite population. The inference for these parameters is based on the probability distribution induced by the sampling design used to get the sample of individuals. In most of the cases for official statistics, the sample design under consideration is complex in the sense of not providing a simple random sample of the population.

After getting a probabilistic sample, sometimes it is necessary to consider the classification of the individuals in the sample through different categories in one or more nominal variables. This classification can be incorporated in a contingency table in order to summarize two variables or the temporal variations in a single variable at two different periods of time. However, in order to get accurate estimates, it is not advisable to ignore the sampling design in the inference for the parameters of interest.

Another common problem in this type of survey is nonresponse for some sample units, which can rarely be considered random or ignorable. Therefore it is necessary to consider some approach that can compensate for the potentially nonignorable nonresponse. Chen and Fienberg (1974), Stasny (1987) and recently Lu and Lohr (2010) have considered two-stage models in order to classify the individuals in a sample for two different times with nonignorable nonresponse. However, this approach ignored the sampling design that is complex and also informative for most surveys conducted for producing official statistics.

This article considers a common scenario for longitudinal surveys where the main aim is to estimate the number of population individuals belonging to several cells in a contingency table according to the categories of a variable measured at two different points in time. We also consider the modeling of the nonresponse that can affect the estimates if it is ignored. The inferential processes are tied to the complex survey design used to collect the information in the sample.

For instance, in labour force surveys, it is possible to find complex classifications depending on the labour force status of the respondents at two consecutive periods of observation and measurement. The

1. Andrés Gutiérrez, Facultad de Estadística. Universidad Santo Tomás. E-mail: hugogutierrez@usantotomas.edu.co; Leonardo Trujillo, Department of Statistics, Universidad Nacional de Colombia. E-mail: ltrujilloo@unal.edu.co; Pedro Luis do Nascimento Silva, Instituto Brasileiro de Geografia e Estatística (IBGE). E-mail: pedro-luis.silva@ibge.gov.br.

aim is to estimate the number of people that in a past period were working and are still working in the current period of observation. Another possible objective is to estimate the number of people who were unemployed in the last period of observation and are still unemployed in the current period of the survey or the number of people that in the last period of observation were employed and in the current period are unemployed or vice versa. For this example, all the entries on Table 1.1 are considered as parameters of interest. Note that even under a census, the counts in Table 1.1 may not be observable due to nonresponse.

Table 1.1
Parameters of interest in a contingency table corresponding to a labour force survey at two consecutive periods of observation.

Period 1	Period 2			
	Employed	**Unemployed**	**Inactive**	**Total**
Employed	X_{11}	X_{12}	X_{13}	X_{1+}
Unemployed	X_{21}	X_{22}	X_{23}	X_{2+}
Inactive	X_{31}	X_{32}	X_{33}	X_{3+}
Total	X_{+1}	X_{+2}	X_{+3}	X_{++}

Kalton (2009) stated that, in terms of the marginal totals, it is possible to estimate the net flows through a direct comparison between the two periods of observation. Then, it is possible to determine if the unemployment rate increased or decreased and also in what magnitude. For example, comparing that on period 1 there were $X_{1+} = \sum_j X_{1j}$ people employed, whereas on period 2 there were $X_{+1} = \sum_i X_{i1}$ people employed. Nevertheless, a more detailed analysis can be obtained analyzing the gross flows as a decomposition of the net flows. In this way, if the unemployment rate increased one percentage point, it is possible to conclude if this increase was due to the fact that one percentage point of the employed people lost their job or because ten percentage points of the employed people lost their job and nine percentage points of the unemployed people found a new job. This is possible comparing the values X_{ij}.

Also, given that in a complex survey it is possible to have unequal sampling weights and clustering and stratification effects, the likelihood function of the sampling data is difficult to find in an analytical way. Then, using classical methods of maximum likelihood would no longer be convenient for survey data from complex surveys. Then, the standard analyses must be modified to take into account the sampling weights and the sampling effects of a complex survey such as weighted estimation of proportions, variance estimation based on the sampling design and generalized corrections for the design effects (Pessoa and Silva 1998).

Section 2 surveys the basic statistical concepts used in this paper, such as survey estimators, nonresponse and categorical data inference. Section 3 proposes a superpopulation model describing the probabilistic behavior of the assignment of the individuals according to the categories of the variable considered in the survey. This corresponds to a two-stage Markov chain model. Some basic concepts of pseudo-likelihood estimation are also reviewed in Section 3. Then, in Section 4, we propose some estimators for the model parameters and the counts in the gross flows contingency table. These estimators are design-unbiased and the mathematical expressions to estimate their variance are shown in Section 5. Section 6 considers both an empirical application and a Monte Carlo simulation in order to test the

proposed methodology when the data in the survey is obtained under a simple and a complex survey design. Our simulation shows that other methodological approaches lead to biased estimation. Section 7 considers a practical application for estimating gross flows for the *Pesquisa Mensal de Emprego* (PME survey) in Brazil. In Section 8, we highlight the strengths and shortcomings of the proposed method. All the mathematical proofs are presented in the Appendix.

2 Motivation

2.1 Sampling designs and estimators

Consider a finite population as a set of N units, where $N < \infty$, forming the universe of study. N is known as the population size. Each element belonging to the population can be identified with an index k. Let U be the index set given by $U = \{1,...,k,...,N\}$. The selection of a sample $s = \{k_1, k_2, \ldots, k_{n(s)}\}$ is done according to a sampling design defined as the multivariate probability distribution over a support Q in a way that $p(s) > 0$ for every $s \in Q$ and

$$\sum_{s \in Q} p(s) = 1.$$

Under a sampling design $p(\cdot)$, an inclusion probability is assigned to every element in the population in order to denote the probability that the element belongs to the sample. For the k-th element in the population this probability is denoted as π_k and it is known as the first order inclusion probability given by

$$\pi_k = Pr(k \in S) = Pr(I_k = 1) = \sum_{s \ni k} p(s)$$

where I_k is a random variable denoting the membership of the element k to the sample, and the subindex $s \ni k$ refers to the sum over all the possible samples containing the k-th element. Analogously, π_{kl} is known as the second order inclusion probability and it denotes the probability that the elements k and l belong to the sample and it is given by

$$\pi_{kl} = Pr(k \in S; l \in S) = Pr(I_k = 1; I_l = 1) = \sum_{s \ni k,l} p(s).$$

The aim of the sample survey is to study a characteristic of interest y associated with every unit in the population and to estimate a function of interest T, called a parameter.

$$T = f(y_1, \ldots, y_k, \ldots, y_N).$$

This inferential approach is known as design-based inference. Under this approach, the estimates of the parameters and their properties depend directly on the discrete probability measure related to the chosen sampling design and do not take into account the properties of the finite population. Also, the values y_k are taken as the observation for the individual k for the characteristic of interest y. Also, y is considered as a fixed quantity rather than a random variable.

Then, the Horvitz-Thompson (HT) estimator can be defined as:

$$\hat{t}_{y,\pi} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k$$

where $d_k = 1/\pi_k$ is the reciprocal of the first-order inclusion probability and it is known as the expansion factor or basic design weight. The HT estimator is unbiased for the total population $t_y = \sum_U y_k$, (assuming all the first order inclusion probabilities are greater than zero) and its variance is given by

$$Var\left(\hat{t}_{y,\pi}\right) = \sum_{k\in U}\sum_{l\in U}\Delta_{kl}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}. \tag{2.1}$$

where $\Delta_{kl} = Cov\left(I_k, I_l\right) = \pi_{kl} - \pi_k\pi_l$. If all the second-order inclusion probabilities are greater than zero, an unbiased estimator of (2.1) is given by

$$\widehat{Var}\left(\hat{t}_{y,\pi}\right) = \sum_{k\in s}\sum_{l\in s}\frac{\Delta_{kl}}{\pi_{kl}}\frac{y_k}{\pi_k}\frac{y_l}{\pi_l}.$$

Gambino and Silva (2009) suggest that in a household survey, the main interest is to focus on characteristics for particular household members that could be related to health variables, educational variables, income/expenses, employment status, etc. In general, the sampling designs used for this kind of survey are complex and use techniques such as stratification, clustering or unequal probabilities of selection. Some of the results from repeated surveys consider the estimation of level at a particular point of time, estimation of changes between two survey rounds and the estimation of the average level parameters over repeated rounds of a survey. Different rotation schemes and the frequency of the survey can affect considerably the precision of the estimators.

2.2 Pseudo-likelihood

Some authors such as Fuller (2009), Chambers and Skinner (2003, p. 179), and Pessoa and Silva (1998, chapter 5) consider the problem where the maximum likelihood estimation is appropriate for simple random samples, as is the case in Stasny (1987), but not for samples resulting from a complex survey design. Under this scheme, it is assumed that the density population function is $f\left(y,\theta\right)$ where the parameter of interest is θ. If there is access to the information for the whole population, through a census, the maximum likelihood estimator of θ can be obtained by maximizing

$$L\left(\theta\right) = \sum_{k\in U}\log f\left(y_k,\theta\right)$$

with respect to θ. We will denote θ_N as the value maximizing the last expression. The likelihood equations for the population are given by

$$\sum_{k\in U}u_k\left(\theta\right) = 0.$$

The u_k are known as *scores* and they are defined as

$$u_k\left(\theta\right) = \frac{\partial\log f\left(y_k,\theta\right)}{\partial\theta}.$$

The pseudo-likelihood approach considers that θ_N is the parameter of interest according to the information collected in a complex sample. If $\sum_{k\in U}u_k\left(\theta\right)$ is considered as the parameter of interest, it is possible to estimate it using a weighted linear estimator

$$\sum_{k\in s} d_k u_k(\theta)$$

where d_k is a sampling design weight such as the inverse of the inclusion probability of the individual k. Then, it is possible to obtain an estimator for θ_N solving the resulting equation system.

Definition 2.1 *A maximum pseudo-likelihood estimator $\hat{\theta}_s$ for θ_N corresponds to the solution of the pseudo-likelihood equations given by*

$$\sum_{k\in s} d_k u_k(\theta) = 0.$$

Using the Taylor linearization method, the asymptotic variance of a maximum pseudo-likelihood estimator based on the sampling design is given by

$$V_p\left(\hat{\theta}_s\right) \approx \left[J\left(\theta_N\right)\right]^{-1} V_p\left[\sum_{k\in s} d_k u_k\left(\theta_N\right)\right]\left[J\left(\theta_N\right)\right]^{-1}$$

where $V_p\left[\sum_{k\in s} d_k u_k\left(\theta_N\right)\right]$ is the variance of the estimator for the population total of the *scores* based on the sampling design and

$$J\left(\theta_N\right) = \left.\frac{\partial \sum_{k\in U} u_k(\theta)}{\partial\theta}\right|_{\theta=\theta_N}.$$

An estimator for $V_p\left(\hat{\theta}_s\right)$ is given by

$$\hat{V}_p\left(\hat{\theta}_s\right) = \left[\hat{J}\left(\hat{\theta}_s\right)\right]^{-1} \hat{V}_p\left[\sum_{k\in s} d_k u_k\left(\hat{\theta}_s\right)\right]\left[\hat{J}\left(\hat{\theta}_s\right)\right]^{-1}$$

where $\hat{V}_p\left[\sum_{k\in s} d_k u_k\left(\hat{\theta}_s\right)\right]$ is a consistent estimator for the variance of the estimator of the population total of the *scores* and

$$\hat{J}\left(\hat{\theta}_s\right) = \left.\frac{\partial \sum_{k\in s} d_k u_k(\theta)}{\partial\theta}\right|_{\theta=\hat{\theta}_s}.$$

Then, following Binder (1983), the asymptotic distribution of $\hat{\theta}_s$ is normal since

$$\hat{V}_p\left(\hat{\theta}_s\right)^{-1/2}\left(\hat{\theta}_s - \theta_N\right) \sim N(0,1).$$

These definitions offer a solid background for the correct inference when using large samples as is the case in labour force surveys.

2.3 Nonresponse

Särndal and Lundström (2005) state that nonresponse has been a topic of increasing interest in national statistical offices during the last decades. Also, in the sampling survey literature, the attention to this topic

has increased considerably. Nonresponse is a common non desirable issue in the development of a survey that can affect considerably the quality of the estimates.

Lohr (1999) discusses several types of nonresponse mechanisms:

- The nonresponse mechanism is ignorable when the probability of an individual responding to the survey does not depend on the characteristic of interest. Note that the word "ignorable" makes reference to a model explaining the mechanism.

- On the other hand, the nonresponse mechanism is nonignorable when the probability of an individual responding to the survey depends on the characteristic of interest. For example, in a labour survey, the possibility of response may depend on the labour force classification of the individuals in a household.

Lumley (2009, chapter 9) analyses individual nonresponse with partial data for a respondent considering a design-based approach adjusting the sampling weights. Fuller (2009, chapter 5) considers some imputation techniques for the nonresponse treatment through probabilistic models and sampling weights. Särndal (2011) considers a model-based approach through balanced sets in order to achieve higher representativeness of the estimates. In the same way, Särndal and Lundström (2010) propose a set of indicators in order to judge the effectiveness of auxiliary information in order to control the bias generated by nonresponse. Särndal and Lundström (2005) give a large number of references about nonresponse. These references examine two main complementary aspects in a survey: prevention of the problem of nonresponse (before it happens) and estimation techniques in order to take into account nonresponse in the inference process. This second aspect is known as adjustment for nonresponse.

3 Markov models for contingency tables with nonresponse

Consider the problem of estimating gross flows between two consecutive periods of time using categorical data obtained from a panel survey and under nonresponse. Also, suppose that the outcome of every interview is the classification of the respondent into any of G possible pairwise disjoint categories, and the aim is to estimate the gross flows between these categories using the information from individuals who were interviewed at two consecutive periods of time. Individuals who either did not answer in one or two periods or were excluded or included for only one of the two periods shall not have a definite classification among the categories. Then, there is one group of individuals with classification between the two periods, a group of individuals who only have the information for one of the two periods and a group of individuals who did not respond in any of the two periods of the survey.

For those individuals responding on times $t-1$ and t, the classification data can be summarized in a matrix of dimension $G \times G$. The available information for those individuals not responding the survey at time $t-1$ but responding at time t can be summarized in a column complement; the information for those individuals not responding at time t but responding at time $t-1$ can be summarized in a row complement. Finally, individuals not responding at any of the two times are included in a single cell counting the number of individuals with missing data at both times.

The whole matrix is illustrated in Table 3.1, where N_{ij} $(i, j = 1,\ldots,G)$ denotes the number of individuals in the population having classification i at time $t-1$ and classification j at time t, R_i denotes the number of individuals not responding at time t and having classification i at time $t-1$, C_j denotes the number of individuals not responding at time $t-1$ and had classification j at time t, and M denotes the number of individuals in the sample not responding in any of the two times. It is important to mention that this analysis does not take into account nonresponse due to the rotation in the survey; it only takes into account individuals belonging to the matched sample ignoring those individuals not responding because they were not selected in the sample.

Table 3.1
Gross flows at two consecutive periods of time.

Time $t-1$			**Time** t		
	1	**2**	$\cdots$	G	**Row complement**
1	N_{11}	N_{12}	$\ldots$	N_{1G}	R_1
2	N_{21}	N_{22}	$\ldots$	N_{2G}	R_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
G	N_{G1}	N_{G2}	$\ldots$	N_{GG}	R_G
Column complement	C_1	C_2	$\ldots$	C_G	M

This paper considers ideas from Stasny (1987) and Chen and Fienberg (1974) - in the sense of considering a maximum likelihood approach in contingency tables for partially classified data - and data resulting from a two-stage process as follows:

1. In the first stage (nonobservable), the individuals are located among the cells of a matrix $G \times G$ according to the probabilities of a Markov chain process. Let η_i be the initial probability of an individual being at the category i at the time $t-1$ with $\sum_i \eta_i = 1$, and p_{ij} be the transition probability from the category i to category j, where $\sum_j p_{ij} = 1$ for every i.

2. In the second stage (observable) of the process, every individual in cell ij can either be nonrespondent at time $t-1$, losing the classification by row; nonrespondent at time t, losing the classification by column; or nonrespondent at both times, losing both classifications.
   - Let ψ be the initial probability of an individual in cell ij responding at time $t-1$.
   - Let ρ_{RR} be the transition probability of classification of the individual in cell ij responding at time $t-1$ and responding at time t.
   - Let ρ_{MM} be the transition probability of an individual in cell ij being a nonrespondent at time $t-1$ and becoming a nonrespondent at time t.

   These probabilities do not depend on the classification stage of the individual.

Data is observed only after the second stage. The aim is to make inferences for the probabilities in the Markov chain process generating the data but also in the chain generating the nonresponse mechanism. In the context of this two-stage model, the corresponding probabilities are shown in Table 3.2.

Table 3.2
Gross flow probabilities at two consecutive times.

Time $t-1$	**Time** t						
	1	**2**	$\cdots$	j	$\cdots$	G	**Row complement**
1							
2							
$\vdots$							
i				$\{\eta_i p_{ij} \psi \rho_{RR}\}$			$\{\sum_j \eta_i p_{ij} \psi (1-\rho_{RR})\}$
$\vdots$							
G							
Column complement				$\{\sum_i \eta_i p_{ij} (1-\psi)(1-\rho_{MM})\}$			$\sum_i \eta_i p_{ij} \sum_j (1-\psi)\rho_{MM}$

In this way, the likelihood function for the observed data under this two-stage model is proportional to

$$\prod_i \prod_j \left[\psi \rho_{RR} \eta_i p_{ij}\right]^{N_{ij}} \times \prod_i \left[\sum_j \psi (1-\rho_{RR}) \eta_i p_{ij}\right]^{R_i} \times \prod_j \left[\sum_i (1-\psi)(1-\rho_{MM}) \eta_i p_{ij}\right]^{C_j} \times \left[\sum_i \sum_j (1-\psi) \rho_{MM} \eta_i p_{ij}\right]^{M}. \tag{3.1}$$

3.1 Parameters of interest

Data are only observed after the second stage and the aim is to make inferences for both the probabilities at the Markov chain generating the data and the chain generating nonresponse. Under this two-stage model, the probabilities of the matrix of data are shown in Table 3.2 and they constitute some of the parameters of interest.

On the other hand, coming from the non-observable process, it is necessary to consider other parameters of interest as follows. Suppose a finite population U exists, having a classification in two periods of time for all its individuals. This is a non-observable process as, even when census data is obtained, it would be not possible to have a complete classification since not all the individuals will be willing to respond. Considering this non-observable process and assuming that there are G possible classifications at each time, the distribution of the gross flows at the population level are shown in Table 3.3.

X_{ij} is the number of units at the finite population with classification i at time $t-1$ and classification j at time t $(i, j = 1, \ldots, G)$. The population size, N, must satisfy the expression:

$$N = \sum_i \sum_j X_{ij}.$$

Table 3.3
Population gross flows (non-observable process) at two consecutive periods of time.

Time $t-1$	Time t					
	1	**2**	$\cdots$	j	$\cdots$	G
1	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1G}
2	X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2G}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{iG}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
G	X_{G1}	X_{G2}	$\cdots$	X_{Gj}	$\cdots$	X_{GG}

Following the non-observable process from the last section, it is supposed that the vector corresponding to the entries at the last contingency table follows a multinomial distribution with a probability vector containing the values $\{\eta_i p_{ij}\}_{i,j=1,\ldots,G}$. This assumes a superpopulation model where the contingency table counts are considered random. In terms of notation, the probability measure considering these counts will be denoted with the subindex ξ. Then, the probability of classification at cell i, j for the k-th individual is

$$\begin{aligned} &P_\xi\left(k \text{ has got a classification } i \text{ at } t\text{-}1 \text{ and classification } j \text{ at } t\right) \\ &= P_\xi\left(k \text{ has got a classification } i \text{ at } t\text{-}1\right) \\ &\times P_\xi\left(k \text{ has got a classification } j \text{ at } t \mid k \text{ has got a classification } i \text{ at } t\text{-}1\right) \\ &= \eta_i p_{ij}. \end{aligned}$$

This treats X_{ij} as a random variable and if the finite population has N individuals, its expected value based on the model is given by

$$E_\xi\left(X_{ij}\right) = N\eta_i p_{ij} = \mu_{ij}. \tag{3.2}$$

Note that this expected value μ_{ij} is one of the most important parameters to be estimated on this paper as it corresponds to the expected value of the gross flows at the population of interest at the two consecutive periods. On the other hand, it is also important to understand that μ_{ij} is a parameter for the two-stage model. Also, the estimators for η_i and p_{ij} are interdependent and determined by the estimations of the defined parameters at the second stage. Let $\boldsymbol{\eta}$ be the vector containing the parameters η_i; and $\mathbf{p}$ be the vector containing the parameters p_{ij}, for every $i, j = 1,\ldots,G$. The final parameters of interest are:

- the model parameters, determined by the vector

$$\boldsymbol{\theta} = \left(\psi', \rho'_{RR}, \rho'_{MM}, \boldsymbol{\eta}', \mathbf{p}'\right)';$$

- the expected value vector of the population counts defined as

$$\boldsymbol{\mu} = \left(\mu_{11}, \ldots, \mu_{ij}, \ldots, \mu_{GG}\right)'.$$

4 Estimation of the parameters of interest

Let N_{ij} be the total number of respondents for the population of interest having a classification i at time $t-1$ and j at time t. Let R_i be the total number of individuals in the population not responding at time t but responding at time $t-1$ with classification i. Let C_j denote the total number of individuals in the population not responding at time $t-1$ but responding at time t with classification j and finally let M be the total number of individuals at the population not responding at any of the two periods of observation. It follows that the total size of the population, N, must satisfy:

$$N = \sum_i \sum_j N_{ij} + \sum_j C_j + \sum_i R_i + M.$$

Defining the following characteristics of interest, it is possible to define the parameters of interest:

$$y_{1ik} = \begin{cases} 1, & \text{if the } k-\text{th individual responds at } t-1 \text{ with classification } i; \\ 0, & \text{otherwise.} \end{cases}$$

$$y_{2jk} = \begin{cases} 1, & \text{if the } k-\text{th individual responds at } t \text{ with classification } j; \\ 0, & \text{otherwise.} \end{cases}$$

Then, the product of these quantities, defined as $y_{1ik}y_{2jk}$, corresponds to a new characteristic of interest taking the value one if the individual has responded at both times and is classified in the cell ij, or zero otherwise. Also,

$$N_{ij} = \sum_{k \in U} y_{1ik} y_{2jk}.$$

Define the following dichotomic characteristics:

$$z_{1k} = \begin{cases} 1, & \text{if the } k-\text{th individual responds at } t-1; \\ 0, & \text{otherwise.} \end{cases}$$

$$z_{2k} = \begin{cases} 1, & \text{if the } k-\text{th individual responds at } t; \\ 0, & \text{otherwise.} \end{cases}$$

It follows that

$$\begin{aligned} R_i &= \sum_{k \in U} y_{1ik}\left(1 - z_{2k}\right) \\ C_j &= \sum_{k \in U} y_{2jk}\left(1 - z_{1k}\right) \\ M &= \sum_{k \in U} \left(1 - z_{1k}\right)\left(1 - z_{2k}\right). \end{aligned}$$

Let w_k denote the weight for the k-th individual corresponding to a specific sampling strategy (sampling design and estimator) in both waves. Then the following expressions represent the estimators of the parameters of interest:

$$\hat{N}_{ij} = \sum_{k\in S} w_k y_{1ik} y_{2jk}$$
$$\hat{R}_i = \sum_{k\in S} w_k y_{1ik}\left(1-z_{2k}\right)$$
$$\hat{C}_j = \sum_{k\in S} w_k y_{2jk}\left(1-z_{1k}\right)$$
$$\hat{M} = \sum_{k\in S} w_k \left(1-z_{1k}\right)\left(1-z_{2k}\right)$$

for N_{ij}, R_i, C_j and M, respectively. Note that an unbiased estimation for the population size is given by

$$\hat{N} = \sum_i \sum_j \hat{N}_{ij} + \sum_j \hat{C}_j + \sum_i \hat{R}_i + \hat{M} = \sum_s w_k v_k$$

where

$$v_k = \sum_i y_{1ik} \sum_j y_{2jk} + \sum_j y_{2jk}\left(1-z_{1k}\right) + \sum_i y_{1ik}\left(1-z_{2k}\right) + \left(1-z_{1k}\right)\left(1-z_{2k}\right).$$

Taking into account the functional form of all the parameters of interest, and noticing that the likelihood function of the model is proportional to (3.1), we arrive at the following result.

Result 4.1 *The log-likelihood for the observed data at the population can be rewritten as*

$$l_U = \sum_{k\in U} f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right) \tag{4.1}$$

where

$$\begin{aligned} & f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right) \\ & = \sum_i \sum_j y_{1ik} y_{2jk} \ln\left(\psi \rho_{RR} \eta_i p_{ij}\right) \\ & + \sum_i y_{1ik}\left(1-z_{2k}\right) \ln\left(\sum_j \psi\left(1-\rho_{RR}\right)\eta_i p_{ij}\right) \\ & + \sum_j y_{2jk}\left(1-z_{1k}\right) \ln\left(\sum_i \left(1-\psi\right)\left(1-\rho_{MM}\right)\eta_i p_{ij}\right) \\ & + \left(1-z_{1k}\right)\left(1-z_{2k}\right) \ln\left(\sum_i \sum_j \left(1-\psi\right)\rho_{MM}\eta_i p_{ij}\right) \end{aligned}$$

where $\mathbf{y}_1$ *is a vector containing the characteristics* y_{1ik}, $\mathbf{y}_2$ *is a vector containing the characteristics* y_{2jk}, $\mathbf{z}_1$ *is a vector containing the characteristics* z_{1k}, *and* $\mathbf{z}_2$ *is a vector containing the characteristics* z_{2k} *(for every* $k = 1,\ldots,N$ *and* $i, j = 1,\ldots,G$*).*

Now, in order to obtain estimators of the parameters, it is necessary to maximize this last function. Using standard techniques of maximum likelihood, the corresponding likelihood equations are given by

$$\sum_{k\in U} \mathbf{u}_k\left(\theta\right) = \mathbf{0}$$

where the vector $\mathbf{u}_k$, commonly known as *scores*, is defined by

$$\mathbf{u}_k(\theta) = \frac{\partial f_k(\theta)}{\partial \theta}.$$

Also, as it is not usual to survey the whole population, a probability sample is selected and the expression $\sum_{k\in U}\mathbf{u}_k(\theta)$ is considered as a population parameter. In this way, considering $w_k = 1/\pi_k$ as the corresponding sampling weights, an unbiased estimator for this sum of *scores* is defined as $\sum_{k\in S} w_k \mathbf{u}_k(\theta)$. The next expression is known as the pseudo-likelihood equation and it is an effective way to find estimators for the model parameters taking into account the sampling weights:

$$\sum_{k\in S} w_k \mathbf{u}_k(\theta) = \mathbf{0}.$$

It is assumed that for the model in this paper, the initial probability of an individual responding at time $t-1$ is the same for all the possible classifications in the survey. Also, the transition probabilities between respondents and nonrespondents do not depend on the classification of the individual in the survey, ρ_{MM} and ρ_{RR}. Considering these assumptions, the following results will let the estimation of the Markov model probabilities take into account the sampling weights.

Result 4.2 *Under the assumptions of the model, the resulting maximum pseudo-likelihood estimators for* ψ, ρ_{RR} *and* ρ_{MM} *are given by*

$$\hat{\psi}_{mpv} = \frac{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j + \hat{M}}$$

$$\hat{\rho}_{RR,mpv} = \frac{\sum_i \sum_j \hat{N}_{ij}}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}$$

$$\hat{\rho}_{MM,mpv} = \frac{\hat{M}}{\sum_j \hat{C}_j + \hat{M}}$$

respectively.

Result 4.3 *Under the assumptions of the model, the resulting maximum pseudo-likelihood estimators for* η_i *and* p_{ij} *are obtained through iteration until convergence of the next expressions*

$$\hat{\eta}_{i,mpv}^{(v+1)} = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)}\right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j}$$

$$\hat{p}_{ij,mpv}^{(v+1)} = \frac{\hat{N}_{ij} + \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)}\right)}{\sum_j \hat{N}_{ij} + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)}\right)}$$

respectively. The superindex (v) *denotes the value of the estimation for the parameters of interest at the* v*-th iteration.*

The results before provide an exhaustive frame for the implementation of the two-stage Markovian model in order to take into account the sampling weights in longitudinal surveys. Another question of interest is how to choose the initial values $\left\{\hat{\eta}_i^{(0)}\right\}$ and $\left\{\hat{p}_{ij}^{(0)}\right\}$. In general, any set of values is valid if they follow the initial restrictions. These are

$$\sum_i \hat{\eta}_i^{(0)} = 1$$

$$\sum_j \hat{p}_{ij}^{(0)} = 1.$$

However, following the guidelines at Chen and Fienberg (1974) and considering the hypothetical case where all of the individuals responded in both periods, then $M = 0, R_i = 0$ (for every $i = 1,\ldots,G$) and $C_j = 0$ (for every $j = 1,\ldots,G$) and their sampling estimations are also null. Given this, and considering the expressions of the resulting estimators, a sensible choice is given by

$$\hat{\eta}_i^{(0)} = \frac{\sum_j \hat{N}_{ij}}{\sum_i \sum_j \hat{N}_{ij}}$$

$$\hat{p}_{ij}^{(0)} = \frac{\hat{N}_{ij}}{\sum_j \hat{N}_{ij}}.$$

Lastly, this iterative approach is commonly implemented for estimation problems by maximum likelihood in contingency tables. However, some approaches for the fit of log-linear models in contingency tables for complex survey designs can be found at Clogg and Eliason (1987), Rao and Thomas (1988), Skinner and Vallet (2010), among others. The next result provides an approach to gross flow estimation considering the sampling weights at both periods of interest.

Result 4.4 *Under the assumptions of the model, a sampling estimator of* μ_{ij} *is*

$$\hat{\mu}_{ij,mpv} = \hat{N}\hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}.$$

5 Properties of the estimators

Following Cassel, Särndal and Wretman (1976), the aim of considering a survey sampling approach is to gather information from just a subset (sample) of units in the finite population that enable us to obtain conclusion for the whole population. During this process, the statistician must face the randomness sources defining the complex stochastic behavior of the inferential process. Although this paper considers the sampling design as the probability measure determining the inference for the parameters and the model, it is necessary to understand that the proposed Markovian model provides another correctly defined measure of probability. Now we obtain some properties of the estimators proposed in the last section.

The aim of this paper is to incorporate the sampling weights in the proposed model and then it is important to get approximately unbiased estimators with respect to the probability measure related to the sampling design for $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$. The following results show some properties of the proposed estimators considered under the complex survey design. In terms of notation, the probability measure induced for the sampling design will be denoted with the subindex p. The following results provide the maximum likelihood estimators for the parameters of interest when instead of getting a sample, the measurement is obtained through a census or complete enumeration of the individuals in the population.

Result 5.1 *Suppose there is complete access to the whole population and the log-likelihood function of the model is given by (4.1), then the maximum likelihood estimators, under the model assumptions are*

$$\psi_U = \frac{\sum_i \sum_j N_{ij} + \sum_i R_i}{\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M}$$

$$\rho_{RR,U} = \frac{\sum_i \sum_j N_{ij}}{\sum_i \sum_j N_{ij} + \sum_i R_i}$$

$$\rho_{MM,U} = \frac{M}{\sum_j C_j + M}$$

$$\eta_{i,U}^{(v+1)} = \frac{\sum_j N_{ij} + R_i + \sum_j \left(C_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j} \tag{5.1}$$

$$p_{ij,U}^{(v+1)} = \frac{N_{ij} + \left(C_j \hat{\eta}_i^{(v)} p_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_j N_{ij} + \sum_j \left(C_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \Big/ \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)} \tag{5.2}$$

where (5.1) and (5.2) must be jointly iterated to convergence.

Result 5.2 *Under the model assumptions, a maximum likelihood estimator of* μ_{ij} *is*

$$\mu_{ij,U} = N \times \eta_{i,U} \times p_{ij,U}$$

where N *corresponds to the population size and* $\eta_{i,U}$ *and* $p_{ij,U}$ *are defined by the last result, respectively.*

Note that both $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ can be defined as descriptive population quantities. Based on the inference approach induced by the maximum likelihood method, there exist estimators $\boldsymbol{\theta}_U = \left(\psi_U', \boldsymbol{\rho}_{RR,U}', \boldsymbol{\rho}_{MM,U}', \boldsymbol{\eta}_U', \mathbf{p}_U' \right)'$ and $\boldsymbol{\mu}_U = \left(\mu_{11,U}, \ldots, \mu_{ij,U}, \ldots, \mu_{GG,U} \right)'$ defined as the corresponding descriptive population quantities making $\boldsymbol{\theta}_{mpv}$ and $\boldsymbol{\mu}_{mpv}$ consistent with regard to the sampling design in

the sense of definition 2 in Pfeffermann (1993). Note also that $\boldsymbol{\theta}_U$ and $\boldsymbol{\mu}_U$ can be calculated only if there is access to the whole finite population.

Following Pessoa and Silva (1998, p. 79), it is possible to assess that under some regularity conditions, it follows that $\boldsymbol{\theta}_U - \boldsymbol{\theta} = o_p(1)$ and $\boldsymbol{\mu}_U - \boldsymbol{\mu} = o_p(1)$. Also, as in many sampling surveys, both the population and the sample size are generally large, then an appropriate estimator of $\boldsymbol{\theta}_U$ is also an appropriate estimator for $\boldsymbol{\theta}$, and an appropriate estimator for $\boldsymbol{\mu}_U$ will be an appropriate estimator for $\boldsymbol{\mu}$.

In the next section, we explore the properties of the estimators proposed above and we discuss about their suitability for our research problem.

5.1 Properties of the count estimators

Result 5.3 *The estimators $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$, $\hat{M}$, and $\hat{N}$ defined in Section 4 are unbiased with regard to the sampling design.*

The proof is quite immediate. The weighting factor w_k corresponds to the inverse of π_k, the inclusion probability associated to the k-th element. All the estimators are of the Horvitz-Thompson class and therefore are unbiased.

Result 5.4 *Making $w_k = 1/\pi_k$, the corresponding variances for $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$, $\hat{M}$ and $\hat{N}$, are given by*

$$
\begin{aligned}
Var_p\left(\hat{N}_{ij}\right) &= \sum_U\sum_U \Delta_{kl}\frac{y_{1ik}y_{2jk}}{\pi_k}\frac{y_{1il}y_{2jl}}{\pi_l} \\
Var_p\left(\hat{R}_i\right) &= \sum_U\sum_U \Delta_{kl}\frac{y_{1ik}\left(1-z_{2k}\right)}{\pi_k}\frac{y_{1il}\left(1-z_{2l}\right)}{\pi_l} \\
Var_p\left(\hat{C}_j\right) &= \sum_U\sum_U \Delta_{kl}\frac{y_{2jk}\left(1-z_{1k}\right)}{\pi_k}\frac{y_{2jl}\left(1-z_{1l}\right)}{\pi_l} \\
Var_p\left(\hat{M}\right) &= \sum_U\sum_U \Delta_{kl}\frac{\left(1-z_{1k}\right)}{\pi_k}\frac{\left(1-z_{1l}\right)}{\pi_l} \\
Var_p\left(\hat{N}\right) &= \sum_U\sum_U \Delta_{kl}\frac{v_k}{\pi_k}\frac{v_l}{\pi_l}.
\end{aligned}
$$

Unbiased estimators for these variances, respectively, are given by

$$
\begin{aligned}
\widehat{Var}_p\left(\hat{N}_{ij}\right) &= \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{y_{1ik}y_{2jk}}{\pi_k}\frac{y_{1il}y_{2jl}}{\pi_l} \\
\widehat{Var}_p\left(\hat{R}_i\right) &= \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{y_{1ik}\left(1-z_{2k}\right)}{\pi_k}\frac{y_{1il}\left(1-z_{2l}\right)}{\pi_l} \\
\widehat{Var}_p\left(\hat{C}_j\right) &= \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{y_{2jk}\left(1-z_{1k}\right)}{\pi_k}\frac{y_{2jl}\left(1-z_{1l}\right)}{\pi_l}
\end{aligned}
$$

$$\widehat{Var}_p\left(\hat{M}\right) = \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{\left(1-z_{1k}\right)}{\pi_k}\frac{\left(1-z_{1l}\right)}{\pi_l}$$

$$\widehat{Var}_p\left(\hat{N}\right) = \sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{v_k}{\pi_k}\frac{v_l}{\pi_l}.$$

On the other hand, if the w_k correspond to calibration weights, then all the estimators considered are asymptotically unbiased and proofs are given in Deville and Särndal (1992). Their corresponding variances are given by Kim and Park (2010).

5.2 Properties of the model probabilities estimators

Result 5.5 *The first-order Taylor approximation for the estimator ψ_{mpv}, defined at the result 4.2 above, around the point $\left(N_{ij}, R_i, C_j, M\right)$ and $i, j = 1,\ldots,G$, is given by the expression*

$$\begin{aligned}\hat{\psi}_{mpv} &\cong \hat{\psi}_0 \\ &= \psi_U + a_1\sum_i\sum_j\left(\hat{N}_{ij} - N_{ij}\right) + a_1\sum_i\left(\hat{R}_i - R_i\right) \\ &\quad + a_2\sum_j\left(\hat{C}_j - C_j\right) + a_2\left(\hat{M} - M\right)\end{aligned}$$

with

$$a_1 = \frac{\sum_j C_j + M}{\left(\sum_i\sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M\right)^2}$$

$$a_2 = -\frac{\sum_i\sum_j N_{ij} + \sum_i R_i}{\left(\sum_i\sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M\right)^2}.$$

Result 5.6 *The first-order Taylor approximation for the estimator $\hat{\rho}_{RR,mpv}$, defined at the result 4.2 above, around the point $\left(N_{ij}, R_i\right)$ and $i, j = 1,\ldots,G$, is given by the expression*

$$\begin{aligned}\hat{\rho}_{RR,mpv} &\cong \hat{\rho}_{RR,0} \\ &= \rho_{RR,U} + a_3\sum_i\sum_j\left(\hat{N}_{ij} - N_{ij}\right) + a_4\sum_i\left(\hat{R}_i - R_i\right)\end{aligned}$$

with

$$a_3 = \frac{\sum_i R_i}{\left(\sum_i\sum_j N_{ij} + \sum_i R_i\right)^2}$$

$$a_4 = -\frac{\sum_i\sum_j N_{ij}}{\left(\sum_i\sum_j N_{ij} + \sum_i R_i\right)^2}.$$

Result 5.7 *The first-order Taylor approximation for the estimator $\hat{\rho}_{MM,mpv}$, defined at the result 4.2 above, around the point $\left(C_j, M\right)$ and $j = 1,\ldots,G,$ is given by the expression*

$$\begin{aligned}\hat{\rho}_{MM,mpv} &\cong \hat{\rho}_{MM,0} \\ &= \rho_{MM,U} + a_5 \sum_j \left(\hat{C}_j - C_j\right) + a_6\left(\hat{M} - M\right)\end{aligned}$$

with

$$a_5 = -\frac{M}{\left(\sum_j C_j + M\right)^2}$$

$$a_6 = -\frac{\sum_j C_j}{\left(\sum_j C_j + M\right)^2}.$$

Result 5.8 *The estimators* $\hat{\psi}_{mpv}$, $\hat{\rho}_{MM,mpv}$ *and* $\hat{\rho}_{RR,mpv}$, *are approximately unbiased for* ψ_U, $\rho_{MM,U}$, $\rho_{RR,U}$.

Result 5.9 *The estimators* $\hat{\eta}_{i,mpv}$ *and* $\hat{p}_{ij,mpv}$, *are approximately unbiased for* $\eta_{i,U}$ *and* $p_{ij,U}$.

Result 5.10 *The approximate variances for the estimators* $\hat{\psi}_{mpv}$, $\hat{\rho}_{MM,mpv}$ *and* $\hat{\rho}_{RR,mpv}$, *are given by*

$$AV_p\left(\hat{\psi}_{mpv}\right) = V_p\left(\sum_s \frac{E_k^{\psi}}{\pi_k}\right) = \sum_U \sum_U \Delta_{kl} \frac{E_k^{\psi}}{\pi_k} \frac{E_l^{\psi}}{\pi_l}$$

$$AV_p\left(\hat{\rho}_{RR,mpv}\right) = V_p\left(\sum_s \frac{E_k^{RR}}{\pi_k}\right) = \sum_U \sum_U \Delta_{kl} \frac{E_k^{RR}}{\pi_k} \frac{E_l^{RR}}{\pi_l}$$

$$AV_p\left(\hat{\rho}_{MM,mpv}\right) = V_p\left(\sum_s \frac{E_k^{MM}}{\pi_k}\right) = \sum_U \sum_U \Delta_{kl} \frac{E_k^{MM}}{\pi_k} \frac{E_l^{MM}}{\pi_l}$$

where

$$\begin{aligned}E_k^{\psi} &= a_1\left(2 - z_{2k}\right) + a_2\left(1 - z_{1k}\right)\left(2 - z_{2k}\right) \\ E_k^{RR} &= a_3 + a_4\left(1 - z_{2k}\right) \\ E_k^{MM} &= a_5\left(1 - z_{1k}\right) + a_6\left(1 - z_{1k}\right)\left(1 - z_{2k}\right).\end{aligned}$$

Result 5.11 *Unbiased estimators for the approximate variances of the estimators* $\hat{\psi}_{mpv}$, $\hat{\rho}_{MM,mpv}$ *and* $\hat{\rho}_{RR,mpv}$, *are given by*

$$\hat{V}\left(\hat{\psi}_{mpv}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k^{\psi}}{\pi_k} \frac{e_l^{\psi}}{\pi_l}$$

$$\hat{V}\left(\hat{\rho}_{RR,mpv}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k^{RR}}{\pi_k} \frac{e_l^{RR}}{\pi_l}$$

$$\hat{V}\left(\hat{\rho}_{MM,mpv}\right) = \sum_s \sum_s \frac{\Delta_{kl}}{\pi_{kl}} \frac{e_k^{MM}}{\pi_k} \frac{e_l^{MM}}{\pi_l}$$

respectively, where

$$
\begin{aligned}
e_k^{\psi} &= \hat{a}_1\left(2-z_{2k}\right)+\hat{a}_2\left(1-z_{1k}\right)\left(2-z_{2k}\right)\\
e_k^{RR} &= \hat{a}_3+\hat{a}_4\left(1-z_{2k}\right)\\
e_k^{MM} &= \hat{a}_5\left(1-z_{1k}\right)+\hat{a}_6\left(1-z_{1k}\right)\left(1-z_{2k}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\hat{a}_1 &= \frac{\sum_j \hat{C}_j+\hat{M}}{\left(\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i+\sum_j \hat{C}_j+\hat{M}\right)^2}\\
\hat{a}_2 &= -\frac{\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i}{\left(\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i+\sum_j \hat{C}_j+\hat{M}\right)^2}\\
\hat{a}_3 &= \frac{\sum_i \hat{R}_i}{\left(\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i\right)^2}\\
\hat{a}_4 &= -\frac{\sum_i\sum_j \hat{N}_{ij}}{\left(\sum_i\sum_j \hat{N}_{ij}+\sum_i \hat{R}_i\right)^2}\\
\hat{a}_5 &= -\frac{\hat{M}}{\left(\sum_j \hat{C}_j+\hat{M}\right)^2}\\
\hat{a}_6 &= -\frac{\sum_j \hat{C}_j}{\left(\sum_j \hat{C}_j+\hat{M}\right)^2}.
\end{aligned}
$$

Result 5.12 *The approximate variances for the estimators* $\hat{\eta}_{i,mpv}$ *and* $\hat{p}_{ij,mpv}$ *are given by*

$$
\begin{aligned}
AV_p\left(\hat{\eta}_{i,mpv}\right) &= \frac{1}{\left(J_{\eta_i}\right)^2}\sum_U\sum_U \Delta_{kl}\frac{u_k\left(\eta_i\right)}{\pi_k}\frac{u_l\left(\eta_i\right)}{\pi_l}\\
AV_p\left(\hat{p}_{ij,mpv}\right) &= \frac{1}{\left(J_{p_{ij}}\right)^2}\sum_U\sum_U \Delta_{kl}\frac{u_k\left(p_{ij}\right)}{\pi_k}\frac{u_l\left(p_{ij}\right)}{\pi_l}
\end{aligned}
$$

where

$$
u_k\left(\eta_i\right) = \frac{\sum_j y_{1ik}y_{2jk}+y_{1ik}\left(1-z_{2k}\right)}{\eta_i}+\sum_j y_{2jk}\left(1-z_{1k}\right)\frac{p_{ij}}{\sum_i \eta_i p_{ij}}+\left(1-z_{1k}\right)\left(1-z_{2k}\right)-1
$$

$$
\begin{aligned}
u_k\left(p_{ij}\right) = \frac{y_{1ik}y_{2jk}}{p_{ij}}+y_{1ik}\left(1-z_{2k}\right)+y_{2jk}\left(1-z_{1k}\right)\frac{\eta_i}{\sum_i \eta_i p_{ij}}+\left(1-z_{1k}\right)\left(1-z_{2k}\right)\eta_i\\
-\frac{1}{\hat{N}}\left(\sum_j \hat{N}_{ij}+\hat{R}_i+\hat{M}\eta_i+\sum_j \hat{C}_j\left(\frac{\eta_i p_{ij}}{\sum_i \eta_i p_{ij}}\right)\right)
\end{aligned}
$$

$$J_{\eta_i} = -\frac{2}{\eta_i^2}\sum_U y_{1ik} + \frac{1}{\eta_i^2}\sum_U y_{1ik} z_{2k} - \sum_U (1-z_{1k})\sum_j \frac{y_{2jk} p_{ij}^2}{\left(\sum_i \eta_i p_{ij}\right)^2}$$

$$J_{p_{ij}} = -\frac{1}{p_{ij}^2}\sum_U y_{1ik} y_{2jk} - \frac{\eta_i^2}{\left(\sum_i \eta_i p_{ij}\right)^2}\sum_U y_{2jk}(1-z_{1k}).$$

Result 5.13 *Unbiased estimators for the approximate variances of the estimators* $\hat{\eta}_{i,mpv}$ *and* $\hat{p}_{ij,mpv}$ *are given by*

$$\hat{V}_p\left(\hat{\eta}_{i,mpv}\right) = \frac{1}{\left(\hat{J}_{\hat{\eta}_i}\right)^2}\sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{\hat{u}_k(\hat{\eta}_i)}{\pi_k}\frac{\hat{u}_l(\hat{\eta}_i)}{\pi_l}$$

$$\hat{V}_p\left(\hat{p}_{ij,mpv}\right) = \frac{1}{\left(\hat{J}_{\hat{p}_{ij}}\right)^2}\sum_s\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\frac{\hat{u}_k(\hat{p}_{ij})}{\pi_k}\frac{\hat{u}_l(\hat{p}_{ij})}{\pi_l}$$

where

$$\hat{u}_k(\hat{\eta}_i) = \frac{\sum_j y_{1ik} y_{2jk} + y_{1ik}(1-z_{2k})}{\hat{\eta}_i} + \sum_j y_{2jk}(1-z_{1k})\frac{\hat{p}_{ij,mpv}}{\sum_i \hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}} + (1-z_{1k})(1-z_{2k})$$

$$\hat{u}_k(\hat{p}_{ij}) = \frac{y_{1ik}y_{2jk}}{\hat{p}_{ij,mpv}} + y_{1ik}(1-z_{2k}) + y_{2jk}(1-z_{1k})\frac{\hat{\eta}_{i,mpv}}{\sum_i \hat{\eta}_{i,mpv} p_{ij,,mpv}} + (1-z_{1k})(1-z_{2k})\hat{\eta}_{i,mpv}$$

and

$$\hat{J}_{\hat{\eta}_i} = -\frac{2}{\hat{\eta}_{i,mpv}^2}\sum_U y_{1ik} + \frac{1}{\hat{\eta}_{i,mpv}^2}\sum_U y_{1ik} z_{2k} - \sum_U (1-z_{1k})\sum_j \frac{y_{2jk}\hat{p}_{ij,mpv}^2}{\left(\sum_i \hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}\right)^2}$$

$$\hat{J}_{\hat{p}_{ij}} = -\frac{1}{\hat{p}_{ij,mpv}^2}\sum_U y_{1ik} y_{2jk} - \frac{\hat{\eta}_{i,mpv}^2}{\left(\sum_i \hat{\eta}_{i,mpv}\hat{p}_{ij,mpv}\right)^2}\sum_U y_{2jk}(1-z_{1k}).$$

5.3 Properties of gross flows estimators

Result 5.14 *Under the model assumptions, the first-order Taylor approximation of the gross flows estimator given by* $\hat{\mu}_{ij}$ *and defined in result 4.4, around the point* $(N, \eta_{i,U}, p_{ij,U})$ *and* $i, j = 1,\ldots,G$, *is given by*

$$\begin{aligned}\hat{\mu}_{ij,mpv} &\cong \hat{\mu}_{ij,0} \\ &= \mu_{ij,U} + a_7\left(\hat{N}_{ij} - N_{ij}\right) + a_8\left(\hat{\eta}_{i,mpv} - \eta_{i,U}\right) + a_9\left(\hat{p}_{ij,mpv} - p_{ij,U}\right)\end{aligned}$$

with

$$a_7 = \eta_{i,U}\, p_{ij,U}$$

$$a_8 = N_{ij}\, p_{ij,U}$$

$$a_9 = N_{ij}\eta_{i,U}.$$

Result 5.15 *The gross flows estimator* $\hat{\mu}_{ij,mpv}$ *is approximately unbiased for* $\mu_{ij,U}$.

Result 5.16 *The following expression approximate the variance for* $\hat{\mu}_{ij,mpv}$

$$AV_p\left(\hat{\mu}_{ij,mpv}\right) \cong a_7^2 Var_p\left(\hat{N}_{ij}\right) + a_8^2 AV_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 AV_p\left(\hat{p}_{ij}\right). \tag{5.3}$$

Result 5.17 *An approximately unbiased estimator for the asymptotic variance in (5.3) is given by*

$$\hat{V}_p\left(\hat{\mu}_{ij,mpv}\right) = \hat{a}_7^2 \hat{V}_p\left(\hat{N}_{ij}\right) + a_8^2 \hat{V}_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 \hat{V}_p\left(\hat{p}_{ij}\right)$$

with

$$\hat{a}_7 = \hat{\eta}_{i,U}\, \hat{p}_{ij,U}$$

$$\hat{a}_8 = \hat{N}_{ij}\, \hat{p}_{ij,U}$$

$$\hat{a}_9 = \hat{N}_{ij}\hat{\eta}_{i,U}.$$

6 Empirical application

We first consider an empirical approach in this section, through simulations that will let us assess some statistical properties such as unbiasedness and efficiency of the proposed estimators. Following the modeling proposed by Stasny (1987), we considered a two-stage simulation as follows:

- Allocation of all the individuals in the population to the different cells of a contingency table. In this first stage, we will define the initial probabilities η_i, p_{ij} and,

- Nonresponse process at two consecutive periods. In this second stage, we will define the initial probabilities ψ, ρ_{RR} and ρ_{MM}.

In the first stage, it was necessary to assume some conditions (non observable process) where the group classification probabilities were established at time $t-1$ and the conditional classification probabilities at time t. In this way, every individual in the population was assumed to be classified in any of three categories: E1, E2 and E3. The state vector at time t was given by

$$\boldsymbol{\eta} = \left(\eta_1, \eta_2, \eta_3\right)' = \left(0.9, 0.05, 0.05\right)'.$$

In this way, there is a classification probability in E1 equals to 0.9 for any individual in the population and classification probabilities in E2 and E3 equal to 0.05. The transition matrix from time $t-1$ to time t is given by

$$\mathbf{P} = \begin{pmatrix} \mathbf{p}_1' \\ \mathbf{p}_2' \\ \mathbf{p}_3' \end{pmatrix} = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} 0.80 & 0.15 & 0.05 \\ 0.30 & 0.60 & 0.10 \\ 0.10 & 0.10 & 0.80 \end{pmatrix}.$$

We assumed that the population size was $N = 100,000$ and its size would not change at the two periods of evaluation. In order to classify the individuals at the periods of time we used the R function rmultinom (R Development Core Team 2012). This way, the distribution of gross flows according to equation (3.2) would be given by the values in Table 6.1.

Table 6.1
Expected values under the model ξ for the population gross flows at two consecutive periods.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	72,000	13,500	4,500
E2	1,500	3,000	500
E3	500	500	4,000

6.1 Methodology

We considered for this empirical exercise, a Monte-Carlo with $L = 1,000$ simulations. In order to classify the individuals between respondents and nonrespondents at the two periods of time, we used the function rmultinom from the language R. Dichotomic variables y_{1ik}, y_{2jk}, z_{1k} and z_{2k} were created using the function Domains from the library TeachingSampling (Gutiérrez 2009).

For each run of the simulation, a sample of size $n = 10,000$ was drawn. We considered a simple random sampling design (SI) along with a complex sampling design inducing unequal inclusion probabilities (π PS). The behavior of the different proposed estimators will be assessed according to their relative bias and relative root mean square error, given by

$$RB = L^{-1}\sum_{l=1}^{L}\frac{\hat{\theta}_l - \theta}{\theta} \qquad \text{and} \qquad RRMSE = \frac{\sqrt{L^{-1}\sum_{l=1}^{L}\left(\hat{\theta}_l - \theta\right)^2}}{\theta}.$$

respectively. In those situations where the vector of inclusion probabilities was unequal, function S.piPS on the library TeachingSampling was used in order to choose a without replacement sample with inclusion probabilities proportional to an auxiliary characteristic assumed known and following a normal distribution with different parameters. The proposed methodology is compared with two other estimators: an estimator taking into account the functional shape of the model but not taking into account the sampling design and a gross flows estimator not taking into account the sampling design but assuming that the nonresponse is ignorable.

The first estimator, that we call the *Design-based estimator*, corresponds to the expressions at results 4.2, 4.3 and 4.4. The second estimator, that we shall call as *Model-based estimator*, correspond to the expressions at result 5.1, being maximum likelihood estimators not considering the sampling weights.

Finally, the third estimator, that we call the *Naive estimator* estimator expands the sampling information to the population and is given by

$$\hat{\mu}_{ij,ING} = \frac{N}{\sum_i \sum_j N_{ij}} N_{ij}.$$

The response probability at time $t-1$ was assumed as $\psi = 0.8$. The response probability at time t for those individuals responding at time $t-1$ was assumed as $\rho_{RR} = 0.9$. Finally, the nonresponse probability at time t for those individuals not responding at time $t-1$ was assumed as $\rho_{MM} = 0.7$.

Based on model ξ, the expected values of the responses are given in Table 6.2.

Table 6.2
Expected values under the model ξ for the response at two consecutive periods.

Time $t-1$	Time t	
	Response	Nonresponse
Response	72,000	8,000
Nonresponse	6,000	14,000

Taking into account the dynamics of the respondents in both periods and assuming that is possible to collect all the population information through a census, we get the classifications given in Table 6.3 below.

Table 6.3
Expected values under the model ξ for the population gross flows (observable process) at two consecutive periods.

Time $t-1$	Time t			
	E1	E2	E3	Row complement
E1	51,840	9,720	3,240	7,200
E2	1,080	2,160	360	400
E3	360	360	2,880	400
Column complement	4,440	1,020	540	14,000

6.2 Results

6.2.1 Simple random sampling: *design-based* and *model-based estimator*

In a first empirical approach, we considered a simple random sampling without replacement as the sampling design. This sampling design induces uniform inclusion probabilities and expansion factors. Under this scenario, the design-based and model-based estimators are the same. Under this scenario the approach shows some strength according to the values of the relative biases that can be considered as negligible. This can be appreciated in Tables 6.4, 6.5 and 6.6.

Table 6.4
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the proposed estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	0.24 (0.094)	-0.35 (0.189)	-0.49 (0.474)
E2	-2.89 (0.158)	-1.89 (0.221)	2.00 (0.980)
E3	-0.63 (0.790)	4.54 (0.822)	-0.84 (0.569)

Table 6.5
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the transition probabilities p_{ij}.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	0.13 (0.284)	-0.39 (0.537)	-1.00 (3.225)
E2	1.70 (1.296)	-2.29 (0.569)	8.64 (0.347)
E3	-6.6 (3.415)	2.09 (1.992)	0.56 (0.158)

Table 6.6
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the initial classification probabilities η_i.

Time $t-1$		
η_1	η_2	η_3
-0.01 (0.094)	-1.42 (0.980)	1.74 (0.790)

Also, the relative bias in percentage for the response probability ψ was -0.23 and the relative root mean square error in percentage was 0.221; for the response probability ρ_{RR}, the bias in percentage was 0.055 and the relative root mean square error in percentage was 0.031; for the nonresponse probability ρ_{MM}, the bias in percentage was -0.192 and and the relative root mean square error in percentage was 0.189. On the other hand, Table 6.7 shows the empirical expected value of the gross flows for the proposed estimator and it can be appreciated that the values are very close to those given on Table 6.1.

Table 6.7
Empirical expected values for the proposed estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	72,085	13,444	4,454
E2	1,504	2,889	535
E3	474	519	4,092

6.2.2 Simple random sampling: *naive estimator*

Under this scenario and considering that this estimator does not take into account the nonresponse process, the values of the relative biases cannot be considered as negligible. This can be appreciated in Table 6.8.

Table 6.8
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the naive estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	-1.21 (4.4)	**10.2** (60.7)	8.34 (25.2)
E2	-0.25 (38.8)	-7.51 (30.9)	1.33 (12.6)
E3	**13.7** (43.3)	-8.54 (46.1)	0.92 (6.9)

Table 6.9 shows the empirical expected values for the naive estimator; compared to the expected values for the model given in Table 6.1 these are not even close.

Table 6.9
Empirical expected values for the naive estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	54,628	760	4,507
E2	1,506	2,079	1,175
E3	1,603	905	32,832

6.2.3 Unequal inclusion probabilities: *design-based estimator*

In a third scenario, we considered a sampling design that induces unequal inclusion probabilities and expansion factors. Under this scenario, the proposed estimators are still unbiased both for the gross flows and for the parameters of the model. The relative biases are shown on Tables 6.10, 6.11 and 6.12.

Table 6.10
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the proposed estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	-0.09 (0.8)	0.25 (3.6)	3.17 (7.9)
E2	0.72 (40.9)	-1.21 (27.2)	-4.62 (71.08)
E3	1.76 (20.4)	-3.19 (22.6)	-0.73 (7.2)

Table 6.11
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the transition probabilities p_{ij}.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	-0.05 (0.7)	0.115 (3.6)	0.47 (7.1)
E2	2.39 (36.0)	-0.13 (18.6)	-6.40 (69.1)
E3	1.15 (24.9)	-5.14 (21.7)	0.49 (3.7)

Table 6.12
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the initial classification probabilities η_i.

Time $t-1$		
η_1	η_2	η_3
-0.02 (1.1)	-0.70 (19.8)	1.13 (6.9)

For the probability of response ψ, the bias in percentage was -0.46 and the relative root mean square error in percentage was 0.6; for the probability of response ρ_{RR}, the bias in percentage was -0.21 and the relative root mean square error in percentage was 0.6; for the probability of nonresponse ρ_{MM}, the bias in percentage was 0.99 and the relative root mean square error in percentage was 1.8. On the other hand, Table 6.13 shows the empirical expected values of the proposed estimator for the population gross flows and these are very close to the values given in Table 6.1.

Table 6.13
Empirical expected values of the design-based estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	71,910	13,505	4,518
E2	1,523	2,972	470
E3	511	479	4,062

6.2.4 Unequal inclusion probabilities: *model-based estimator*

A fourth scenario considers a sampling design inducing unequal inclusion probabilities and expansion factors in the same way as the last scenario. However, we consider estimators not taking into account the sampling design only the model ξ. Under this scenario, estimations are biased for both the gross flows and the model parameters as can be appreciated by the relative biases on Tables 6.14, 6.15 and 6.16.

Table 6.14
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the model-based estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	4.7 (6.1)	4.6 (8.9)	6.3 (10.5)
E2	**-89.0** (126.6)	**-89.5** (125.9)	**-88.4** (126.9)
E3	4.1 (23.8)	-3.7 (26.67)	5.3 (10.4)

Table 6.15
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the transition probabilities p_{ij}.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	0.03 (0.9)	-0.71 (4.1)	1.63 (8.6)
E2	2.77 (35.5)	-1.50 (19.6)	0.70 (70.6)
E3	4.00 (20.8)	**-14.6** (20.1)	1.33 (3.41)

Table 6.16
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) of the initial probabilities of classification η_i.

Time $t-1$		
η_1	η_2	η_3
4.74 (6.48)	**-89.3** (126.7)	3.95 (11.9)

In the same way, for the response probability ψ, the relative bias in percentage was -0.77 and the relative root mean square error was 1.7; for the response probability ρ_{RR}, the relative bias in percentage was -0.53 and the relative root mean square error was 0.5; for the nonresponse probability ρ_{MM}, the relative bias in percentage was 0.11 and the relative root mean square error was 1.8. On the other hand, Table 6.17 shows the empirical expected values for the model-based estimator for the population gross flows (not considering the sampling design) and these are quite far from the values in Table 6.1, especially for the second category.

Table 6.17
Empirical expected values for the model-based estimator for the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	75,438	14,039	4,790
E2	164	315	53
E3	540	443	4,213

6.2.5 Unequal inclusion probabilities: *naive estimator*

In a fifth scenario, we consider a sampling design with unequal inclusion probabilities and expansion factors. Considering the naive estimator, that does not take into account the sampling design nor the respondent model, Table 6.18 shows the relative bias for each cell in the matrix of gross flows. This estimator would be only recommendable if the nonresponse was ignorable and the sampling design would correspond to a simple random sampling design.

Table 6.18
Relative biases in percentage and relative root mean square errors in percentage (shown in brackets) for the naive estimator of the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	-28.1 (34.7)	-27.6 (60.0)	-24.5 (41.1)
E2	**497.0** (629.2)	**570.2** (610.3)	**432.7** (686.2)
E3	**-40.5** (44.2)	**-37.0** (47.4)	**-33.0** (33.8)

In order to have a more accurate comparison, it would be possible to calculate the expected values of the gross flows and compare them with the current scenario. Table 6.19 shows the empirical expected values for the naive estimator; compared to the expected values for the model given in Table 6.1 these are not very close and are especially poor for the classifications in the second category.

Table 6.19
Empirical expected values for the naive estimator of the population gross flows.

Time $t-1$	Time t		
	E1	**E2**	**E3**
E1	51,755	9,849	3,297
E2	9,194	19,838	2,823
E3	279	295	2,665

7 Actual application: estimation of population gross flows for the PME survey

The *Pesquisa Mensal de Emprego* (PME - Brazilian Monthly Labour Survey) is a survey providing monthly indicators about the labour market in the main metropolitan areas in Brazil. Its main aim is to estimate the monthly work force and to evaluate the fluctuations and tendencies of the metropolitan labour market. It is also possible to get indicators regarding the effects of the economic conditions in the labour market and to satisfy important needs for policy planning and socio-economic development. This survey has been conducted since 1980, with some major methodological changes in 1982, 1988, 1993 and 2001 (IBGE 2007).

This section illustrates the use of the proposed estimators and the final results for the PME are shown. We will consider the panel P6 from this survey from November, 2010 to February, 2011 and then from November, 2011 to February, 2012. This window of observation administered 21,374 interviews to different people. We have chosen the first two measurements of the panel (November and December, 2010) in order to implement the proposed estimation procedure for the corresponding gross flows. Following an algorithm using the library TeachingSampling (Gutiérrez 2009), we obtain the classification at panel P6, for the months of November and December, 2010 given in Table 7.1.

Table 7.1
Labour classification and response for the occupation level in the sample of panel P6 of the PME survey.

November 2010	December 2010				
	Employed	Unemployed	Inactive	Not in the labour force	Row complement
Employed	5,231	62	227	10	386
Unemployed	51	183	113	0	28
Inactive	235	93	4,200	12	281
Not in the labour force	2	0	17	1,426	96
Column complement	499	27	372	132	7,691

However, since panel P6 corresponds to a probabilistic complex sample of the metropolitan areas in Brazil, every individual in the panel represents themselves and other additional people in the population. Then, using the proposed estimation procedure in this paper and using the corresponding expansion factors from the survey, we notice that the estimated population values for panel P6 correspond to those obtained in Table 7.2.

Table 7.2
Estimated contingency table for the population showing level of occupation and nonresponse at the two considered measurements for the panel P6 in the PME survey.

November 2010	December 2010				
	Employed	Unemployed	Inactive	Not in the labour force	Row complement
Employed	2,162,635	20,602	76,303	3,074	160,768
Unemployed	16,233	80,169	37,786	0	11,504
Inactive	70,551	31,822	1,707,675	6,018	122,412
Not in the labour force	958	0	7,035	566,530	38,171
Column complement	205,033	9,293	136,146	53,640	3,076,388

Using the estimation procedure proposed in this paper, we computed the estimated population gross flows given in Table 7.3. The corresponding estimators are unbiased under the complex design of the PME survey. According to this, the number of employed people in both periods of measurement is estimated as 3,913,274, whereas the number of inactive people for both periods is estimated as 3,035,463.

Table 7.3
Population estimated gross flows for both periods at the PME survey. Estimated coefficients of variation in percentage are shown in brackets.

November 2010	December 2010			
	Employed	Unemployed	Inactive	Not in the labour force
Employed	3,913,274 (0.2)	36,570 (3.1)	136,102 (1.6)	5,573 (7.2)
Unemployed	29,776 (3.5)	144,253 (1.7)	68,320 (2.1)	0 (-)
Inactive	127,193 (1.6)	56,296 (2.3)	3,035,463 (0.3)	10,872 (6.5)
Not in the labour force	1,727 (17.3)	0 (-)	12,496 (5.8)	1,022,836 (0.5)

The estimates in the last table above are the result of the proposed estimation procedure in this paper. Next, we show the estimated parameters on the first stage of the model, defined as the transition probabilities from one category to another in both observation periods.

Table 7.4
Estimation of the probabilities p_{ij}. Estimated coefficients of variation in percentage are shown in brackets.

November 2010	December 2010			
	Employed	Unemployed	Inactive	Not in the labour force
Employed	0.9564 (0.1)	0.0089 (3.1)	0.0332 (1.6)	0.0013 (7.2)
Unemployed	0.1228 (3.4)	0.5952 (1.1)	0.2819 (2.0)	0 (-)
Inactive	0.0393 (1.5)	0.0174 (2.3)	0.9398 (0.1)	0.0033 (6.5)
Not in the labour force	0.0016 (17.6)	0 (-)	0.0120 (5.8)	0.9862 (0.1)

The initial probabilities of classification on the first period of interest are shown in Table 7.5. It can be noticed that, for this particular survey, the highest classification probabilities can be found for the categories of employed and inactive.

Table 7.5
Estimation of the probabilities η_i. Estimated coefficients of variation in percentage are shown in brackets.

November 2010			
η_1	η_2	η_3	η_4
0.4757 (0.2)	0.0281 (1.2)	0.3755 (0.3)	0.1205 (0.5)

Finally, the general response probability was estimated as $\hat{\psi}_{mpv} = 0.595$ (with an estimated coefficient of variation of 0.1%). That means that the rate of response is around 60%. Also, the transition probability that a nonrespondent in the first period is still a nonrespondent the next time was estimated as $\hat{\rho}_{MM,mpv} = 0.883$ (with an estimated coefficient of variation of 0.1%). The transition probability that a respondent in the first period stays on as a respondent the next time was estimated as $\hat{\rho}_{RR,mpv} = 0.934$ (with an estimated coefficient of variation of 0.1%). In general terms, it is possible to state that a status response of an individual in the first period is not changing significatively by the second.

8 Conclusions

This paper has considered a common problem in survey sampling applications. Using superpopulation Markov chain based models, a new methodology was proposed leading to approximately unbiased estimators of gross flows at different times for the particular case of data coming from complex surveys with unequal sampling weights. Possible applications of the methodology in this paper are broad in the case of, for example, national statistical offices considering complex surveys. Life quality or labour force surveys are usually concerned about the estimation of gross flows. However, the possible extensions of this methodology could be applied to the public policy sector for impact evaluations having a classification of the respondents before and after an intervention.

Also we present a solution to a general problem such as nonignorable nonresponse. Models where the nonresponse is not differentiated at different periods or by classification status were considered. However, in some practical applications, it is possible that this is not the case.

The approach of this paper considers that design weights for units between the two time periods are the same. Further work will try to consider different weights between waves by considering either a two-phase sampling scheme or a calibration approach in two-stages. Indeed, it would be of interest to compare the performance of the methodology given in this paper with the calibration methodology. One could consider the approach of Ash (2005) and Sikkel, Hox and de Leeuw (2008) to calibrate in two periods along with Särndal and Lundström (2005) for handling nonresponse.

Further work will try to extend this methodology for more complex Markov chain models in order to consider different sampling weights. A new definition of parameters in the model will be necessary. Also, this methodology could be extended to the case of gross flows in more than two periods of time where classification errors are taken into account.

Acknowledgements

The authors wish to thank two anonymous referees for their constructive comments on an earlier version of this manuscript which resulted in this improved version. Also, the first author wishes to thank Universidad Santo Tomas for the financial support during his PhD studies. This paper is a result of the PhD thesis of Andrés Gutiérrez at Universidad Nacional de Colombia under supervision of the other two authors.

Appendix

A.1 Mathematical proofs of the results on the paper

In this section, the mathematical proofs for some of the most important results in this paper are included.

Proof of Result 4.1

Proof. Taking logarithm to the likelihood function, and defining it as l, it follows that

$$
\begin{aligned}
l_U &= \ln\left(L_U\right) \\
&= \sum_i \sum_j N_{ij} \ln\left(\psi \rho_{RR} \eta_i p_{ij}\right) + \sum_i R_i \ln\left(\sum_j \psi\left(1-\rho_{RR}\right)\eta_i p_{ij}\right) \\
&\quad + \sum_j C_j \ln\left(\sum_i (1-\psi)(1-\rho_{MM})\eta_i p_{ij}\right) + M \ln\left(\sum_i \sum_j (1-\psi)\rho_{MM}\eta_i p_{ij}\right).
\end{aligned}
$$

Note that $N_{ij} = \sum_{k\in U} y_{1ik} y_{2jk}$, $R_i = \sum_{k\in U} y_{1ik}\left(1-z_{2k}\right)$, $C_j = \sum_{k\in U} y_{2jk}\left(1-z_{1k}\right)$ and $M = \sum_{k\in U}\left(1-z_{1k}\right)\left(1-z_{2k}\right)$. After factorizing the sum over the whole population, the result is finally obtained.

Proof of Result 4.2

Proof. Starting from the definition of pseudo-likelihood and taking into account the model assumptions, it follows that

$$
\begin{aligned}
l_U &= \sum_{k\in U}\Bigg[\sum_i \sum_j y_{1ik} y_{2jk}\left[\ln(\psi) + \ln\left(\rho_{RR}\right) + \ln\left(\eta_i\right) + \ln\left(p_{ij}\right)\right] \\
&\quad + \sum_i y_{1ik}\left(1-z_{2k}\right)\left[\ln(\psi) + \ln\left(1-\rho_{RR}\right) + \ln\left(\eta_i\right) + \ln\left(\sum_j p_{ij}\right)\right] \\
&\quad + \sum_j y_{2jk}\left(1-z_{1k}\right)\left[\ln(1-\rho_{MM}) + \ln(1-\psi) + \ln\left(\sum_i \eta_i p_{ij}\right)\right] \\
&\quad + \left(1-z_{1k}\right)\left(1-z_{2k}\right)\left[\ln(1-\psi) + \ln\left(\rho_{MM}\right) + \ln\left(\sum_i \sum_j \eta_i p_{ij}\right)\right]\Bigg] \\
&= \sum_{k\in U} f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right).
\end{aligned}
$$

The *score* for ψ can be defined as

$$
\begin{aligned}
u_k(\psi) &= \frac{\partial f_k\left(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2\right)}{\partial \psi} \\
&= \frac{(1-\psi)\left(\sum_i \sum_j y_{1ik} y_{2jk} + \sum_i y_{1ik}\left(1-z_{2k}\right)\right) - \psi\left(\sum_j y_{2jk}\left(1-z_{1k}\right) + \left(1-z_{1k}\right)\left(1-z_{2k}\right)\right)}{\psi(1-\psi)}.
\end{aligned}
$$

Then, for this parameter, the pseudo-likelihood equations are given by

$$
\sum_{k\in S} w_k u_k(\psi) = 0.
$$

Solving for ψ, it is obtained that

$$
\hat{\psi}_{mpv} = \frac{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_{i\,i} + \sum_j \hat{C}_j + \hat{M}}.
$$

Following an analogous process for the remaining parameters, the result is obtained.

Proof of Result 4.3

Proof. First, it is necessary to warn that the estimation for these parameters is subject to the restrictions $\sum_i \eta_i = 1$ and $\sum_j p_{ij} = 1$. Then, the process must consider the use of Lagrange multipliers. The function to be maximized, including these restrictions, can be expressed as

$$l_U + \lambda_1\left(\sum_i \eta_i - 1\right) + \lambda_2\left(\sum_j p_{ij} - 1\right).$$

Then, the corresponding *score* for η_i is defined by

$$\begin{aligned} u_k(\eta_i) &= \frac{\partial f_k(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2)}{\partial \eta_i} + \frac{\partial \lambda_1\left(\sum_i \eta_i - 1\right)}{\partial \eta_i} \\ &= \frac{\sum_j y_{1ik} y_{2jk} + y_{1ik}(1 - z_{2k})}{\eta_i} + \sum_j y_{2jk}(1 - z_{1k})\frac{p_{ij}}{\sum_i \eta_i p_{ij}} + (1 - z_{1k})(1 - z_{2k}) + \lambda_1. \end{aligned}$$

The last step takes into account the restrictions, since $\sum_i \sum_j \eta_i p_{ij} = \sum_i \eta_i \sum_j p_{ij} = \sum_i \eta_i = 1$. Then, for this parameter, the pseudo-likelihood equations are given by

$$\sum_{k \in S} w_k u_k(\eta_i) = 0.$$

Then, after some algebra, it follows that

$$\eta_i = \frac{\sum_j \sum_s w_k y_{1ik} y_{2jk} + \sum_s w_k y_{1ik}(1 - z_{2k}) + \sum_j \sum_s w_k y_{2jk}(1 - z_{1k})\left(\eta_i p_{ij} / \sum_i \eta_i p_{ij}\right)}{-\sum_s w_k (1 - z_{1k})(1 - z_{2k}) - \lambda_1 \sum_s w_k}.$$

Besides, using the restriction $\sum_i \eta_i = 1$ and adding up over i, it follows that

$$\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j = \left(-\sum_s w_k (1 - z_{1k})(1 - z_{2k}) - \lambda_1 \sum_s w_k\right).$$

Then, we finally obtain that

$$\eta_i = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\hat{C}_j \eta_i p_{ij} / \sum_i \eta_i p_{ij}\right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j}.$$

On the other hand, in order to find the maximum pseudo-likelihood estimator of $\{p_{ij}\}$, the *score* for p_{ij} is defined as

$$\begin{aligned} u_k(p_{ij}) &= \frac{\partial f_k(\psi, \rho_{RR}, \rho_{MM}, \boldsymbol{\eta}, \mathbf{p}, \mathbf{y}_1, \mathbf{y}_2, \mathbf{z}_1, \mathbf{z}_2)}{\partial p_{ij}} + \frac{\partial \lambda_2\left(\sum_i p_{ij} - 1\right)}{\partial p_{ij}} \\ &= \frac{y_{1ik} y_{2jk}}{p_{ij}} + y_{1ik}(1 - z_{2k}) + y_{2jk}(1 - z_{1k})\frac{\eta_i}{\sum_i \eta_i p_{ij}} + (1 - z_{1k})(1 - z_{2k})\eta_i + \lambda_2. \end{aligned}$$

Hence,

$$p_{ij} = \frac{\sum_s w_k y_{1ik} y_{2jk} + \sum_s w_k y_{2jk} (1 - z_{1k}) p_{ij} \eta_i / \sum_i \eta_i p_{ij}}{-\sum_s w_k y_{1ik} (1 - z_{2k}) - \sum_s w_k (1 - z_{1k})(1 - z_{2k}) \eta_i - \sum_s w_k \lambda_2}.$$

Using the restriction $\sum_j p_{ij} = 1$ and adding up over j on both sides, it follows that

$$\sum_j \hat{N}_{ij} + \sum_j \hat{C}_j \frac{p_{ij} \eta_i}{\sum_i \eta_i p_{ij}} = \left(-\sum_s w_k y_{1ik} (1 - z_{2k}) - \sum_s w_k (1 - z_{1k})(1 - z_{2k}) \eta_i - \sum_s w_k \lambda_2 \right).$$

Then, it follows that

$$p_{ij} = \frac{\hat{N}_{ij} + \left(\hat{C}_j \eta_i p_{ij} / \sum_i \eta_i p_{ij} \right)}{\sum_j \hat{N}_{ij} + \sum_j \left(\hat{C}_j \eta_i p_{ij} / \sum_i \eta_i p_{ij} \right)}.$$

Now, note that it is impossible to solve the last expression for the $\{p_{ij}\}$ in such a way that the solution is a closed expression. The same happens with the expression for the $\{\eta_i\}$. However, it is possible to use an iterative approach, which has proven to have a fast convergence in maximum likelihood estimation problems for contingency tables. This approach assumes that the maximum pseudo-likelihood estimator can be found after jointly iterating the following expressions at step $(v+1)$, for $v \geq 1$,

$$\hat{\eta}_{i,mpv}^{(v+1)} = \frac{\sum_j \hat{N}_{ij} + \hat{R}_i + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} / \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_i \sum_j \hat{N}_{ij} + \sum_i \hat{R}_i + \sum_j \hat{C}_j}$$

$$\hat{p}_{ij,mpv}^{(v+1)} = \frac{\hat{N}_{ij} + \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} / \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}{\sum_j \hat{N}_{ij} + \sum_j \left(\hat{C}_j \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} / \sum_i \hat{\eta}_i^{(v)} \hat{p}_{ij}^{(v)} \right)}.$$

This particular iterative procedure was used initially for the formulation of nested likelihood models by Hocking and Oxspring (1971). However, it also appears implemented by Blumenthal (1968), Reinfurt (1970), Chen and Fienberg (1974), Fienberg and Stasny (1983), Stasny (1987), Stasny (1988), and others.

Proof of Result 5.5

Proof. The non-linear estimator $\hat{\psi}_{mpv}$, can be expressed as a function of the estimated totals $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$ and $\hat{M}$ (with $i, j = 1, \ldots, G$). Then,

$$\hat{\psi}_{mpv} = f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right).$$

Finally, the first order Taylor approximation at the point $\left(\hat{N}_{ij} = N_{ij}, \hat{R}_i = R_i, \hat{C}_j = C_j, \hat{M} = M \right)$ is given by

$$\hat{\psi}_{mpv} = \psi_U + a_1 \sum_i \sum_j \left(\hat{N}_{ij} - N_{ij} \right) + a_1 \sum_i \left(\hat{R}_i - R_i \right) + a_2 \sum_j \left(\hat{C}_j - C_j \right) + a_2 \left(\hat{M} - M \right)$$

with

$$a_1 = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{R}_i} \right|_{\substack{\hat{N}_{ij}=N_{ij} \\ \hat{R}_i=R_i \\ \hat{C}_j=C_j \\ \hat{M}=M}} = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{N}_{ij}} \right|_{\substack{\hat{N}_{ij}=N_{ij} \\ \hat{R}_i=R_i \\ \hat{C}_j=C_j \\ \hat{M}=M}} = \frac{\sum_j C_j + M}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M \right)^2}$$

and

$$a_2 = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{C}_j} \right|_{\substack{\hat{N}_{ij}=N_{ij} \\ \hat{R}_i=R_i \\ \hat{C}_j=C_j \\ \hat{M}=M}} = \left. \frac{\partial f\left(\hat{N}_{ij}, \hat{R}_i, \hat{C}_j, \hat{M} \right)}{\partial \hat{M}} \right|_{\substack{\hat{N}_{ij}=N_{ij} \\ \hat{R}_i=R_i \\ \hat{C}_j=C_j \\ \hat{M}=M}} = -\frac{\sum_i \sum_j N_{ij} + \sum_i R_i}{\left(\sum_i \sum_j N_{ij} + \sum_i R_i + \sum_j C_j + M \right)^2}.$$

Proof of Result 5.8

Proof. Calculating the expected value under the sampling design, it follows that

$$\begin{aligned} AE_p\left(\hat{\psi}_{mpv} \right) &\cong E_p\left(\hat{\psi}_0 \right) \\ &= \psi_U + a_1 \sum_i \sum_j \left(E_p\left(\hat{N}_{ij} \right) - N_{ij} \right) + a_1 \sum_i \left(E_p\left(\hat{R}_i \right) - R_i \right) \\ &\quad + a_2 \sum_j \left(E_p\left(\hat{C}_j \right) - C_j \right) + a_2 \left(E_p\left(\hat{M} \right) - M \right) \\ &= \psi_U . \end{aligned}$$

Following a similar process for the remaining estimators, the result is obtained. This proof is a result of the application of the pseudo-likelihood method that induces unbiased estimations for the population parameters in the model as it is proved on Corollary 1 at Binder (1983, p. 291).

Proof of Result 5.10

Proof. Considering $\hat{\psi}_{mpv}$, replacing the expressions for $\hat{N}_{ij}$, $\hat{R}_i$, $\hat{C}_j$, $\hat{M}$ and after some algebraic simplifications, the approximate variance can be expressed as

$$AV\left(\hat{\psi}_{mpv} \right) = Var\left(a_1 \sum_i \sum_j \hat{N}_{ij} + a_1 \sum_i \hat{R}_i + a_2 \sum_j \hat{C}_j + a_2 \hat{M} \right) = Var\left(\sum_{k \in S} \frac{E_k^{\psi}}{\pi_k} \right).$$

Initially, we have that

$$E_k^{\psi} = a_1 \sum_i \sum_j y_{1ik} y_{2jk} + a_1 \sum_i y_{1ik} \left(1 - z_{2k} \right) + a_2 \sum_j y_{2jk} \left(1 - z_{1k} \right) + a_2 \left(1 - z_{1k} \right)\left(1 - z_{2k} \right).$$

Then, using that $\sum_i \sum_j y_{1ik} y_{2jk} = \sum_i y_{1ik} = \sum_j y_{2jk} = 1$ and after some algebra, it follows that

$$E_k^{\psi} = a_1\left(2 - z_{2k}\right) + a_2\left(1 - z_{1k}\right)\left(2 - z_{2k}\right).$$

After an analogous process for $\hat{\rho}_{RR,mpv}$ and $\hat{\rho}_{MM,mpv}$, the variance expressions at the heading of this result are obtained.

Proof of Result 5.12

Proof. The proof is obtained following expression (3.3) at Binder (1983), and taking into account that

$$J_{\eta_i} = \frac{\partial \sum_U u_k\left(\eta_i\right)}{\partial \eta_i}$$

$$J_{p_{ij}} = \frac{\partial \sum_U u_k\left(p_{ij}\right)}{\partial p_{ij}}.$$

Also,

$$\frac{\partial u_k\left(\eta_i\right)}{\partial \eta_i} = -\frac{2y_{1ik} - y_{1ik}z_{2k}}{\eta_i^2} - \left(1 - z_{1k}\right)\sum_j \frac{y_{2jk}p_{ij}^2}{\left(\sum_i \eta_i p_{ij}\right)^2}$$

$$\frac{\partial u_k\left(p_{ij}\right)}{\partial p_{ij}} = -\frac{y_{1ik}y_{2jk}}{p_{ij}^2} - \frac{\eta_i^2}{\left(\sum_i \eta_i p_{ij}\right)^2} y_{2jk}\left(1 - z_{1k}\right).$$

Proof of Result 5.16

Proof.

$$\begin{aligned} AV_p\left(\hat{\mu}_{ij,mpv}\right) &= a_7^2 Var_p\left(\hat{N}_{ij}\right) + a_8^2 AV_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 AV_p\left(\hat{p}_{ij}\right) \\ &\quad + 2a_7a_8 Cov\left(\hat{N}_{ij}, \hat{\eta}_{i,mpv}\right) + 2a_7a_9 Cov\left(\hat{N}_{ij}, \hat{p}_{ij}\right) 2a_8a_9 Cov\left(\hat{\eta}_{i,mpv}, \hat{p}_{ij}\right) \\ &\cong a_7^2 Var_p\left(\hat{N}_{ij}\right) + a_8^2 AV_p\left(\hat{\eta}_{i,mpv}\right) + a_9^2 AV_p\left(\hat{p}_{ij}\right). \end{aligned}$$

This due to

$$\begin{aligned} Cov\left(\hat{N}_{ij}, \hat{\eta}_{i,mpv}\right) &= E_p\left(\hat{N}_{ij}\hat{\eta}_{i,mpv}\right) - E_p\left(\hat{N}_{ij}\right)E_p\left(\hat{\eta}_{i,mpv}\right) \\ &\cong \hat{N}_{ij,U}\eta_{i,U} - \hat{N}_{ij,U}\eta_{i,U} = 0. \end{aligned}$$

Then, it is possible to get:

$$E_p\left(\hat{N}_{ij}\hat{\eta}_{i,mpv}\right) \cong \hat{N}_{ij,U}, \eta_{i,U}$$

using Taylor linearization for $\left(\hat{N}_{ij,U}, \eta_{i,U}\right)$. The other covariances are obtained in a similar way.

References

Ash, S. (2005). Calibration weights for estimators of longitudinal data with an application to the National Long Term Care Survey. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*. American Statistical Association: Alexandria, VA, 2694–2699.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, 63, 542-551.

Cassel, C.M., Särndal, C.E. and Wretman, J.H. (1976). *Foundations of Inference in Survey Sampling*. New York: Wiley.

Chambers, R.L. and Skinner, C.J. (2003). *Analysis of Survey Data*. John Wiley and Sons, Chichester: UK.

Chen, T. and Fienberg, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 30, 629-642.

Clogg, C.C. and Eliason, S.R. (1987). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16, 8-44.

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

Fienberg, S.E. and Stasny, E.A. (1983). Estimating monthly gross flows in labour force participation. *Survey Methodology*, 9(1), 77-102.

Fuller, W.A. (2009). *Sampling Statistics*. Wiley.

Gambino, J.G. and Silva, P.L. (2009). Sampling and estimation in household surveys. In D. Pfeffermann and C.R. Rao (Eds.), *Handbook of Statistics*. Vol. 29A. Sample Surveys: Design, Methods and Applications (pp. 407-439). Amsterdam: Elsevier.

Gutiérrez, H.A. (2009). TeachingSampling: Sampling designs and parameter estimation in finite population. R package version 2.0.1.

Hocking, R.R. and Oxspring, H.H. (1971). Maximun likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, 66, 65-70.

IBGE (2007). *Pesquisa Mensal de Emprego*. Vol. 23, 2nd edition.

Kalton, G. (2009). Designs for surveys over time. In D. Pfeffermann and C.R. Rao (Eds.), *Handbook of Statistics*. Vol. 29A. Sample Surveys: Design, Methods and Applications (pp. 89-108). Amsterdam: Elsevier.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78, 21-39.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxbury Press.

Lu, Y. and Lohr, S. (2010). Gross flow estimation in dual frame surveys. *Survey Methodology*, 36(1), 13-22.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R*. New York: Wiley.

Pessoa, D.G.C. and Silva, P.L. (1998). *Análise de Dados Amostrais Complexos*. São Paulo : Associação Brasileira de Estatística.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

R Development Core Team (2012). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0.

Rao, J.N.K. and Thomas, D.R. (1988). The analysis of cross-classified data from complex surveys. *Sociological Methodology*, 18, 213-269.

Reinfurt, D.W. (1970). The analyis of categorical data with supplemented margins including applications to mixed models. Unpublished Ph.D dissertation. Department of Biostatistics. University of North Carolina.

Särndal, C.E. (2011). The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection. *Journal of Official Statistics*, 27, 1-21.

Särndal, C.E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley and Sons, Chichester: UK.

Särndal, C.E. and Lundström, S. (2010). Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36(2), 131-144.

Sikkel, D., Hox, J. and de Leeuw, E. (2008). Using auxiliary data for adjustment in longitudinal research. In P. Lynn (Ed), *Methodology of longitudinal surveys*. New York: Wiley. An earlier version is available at http://www.iser.essex.ac.uk/ulsc/mols2006/programme/data/papers/Sikkel.pdf

Skinner, C.J. and Vallet, L.A. (2010). Fitting log-linear models to contingency tables from surveys with complex sampling designs: An investigation of the Clogg-Eliason approach. *Sociological Methods and Research*, 39, 83-108.

Stasny, E.A. (1987). Some Markov-chain models for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 3, 359-373.

Stasny, E.A. (1988). Modeling nonignorable nonresponse in categorical panel data with an example in estimating gross labor-flows. *Journal of Business and Economic Statistics*, 6, 207-219.