Improving The Quality Of Survey Estimates From Longitudinal Studies

Piero Demetrio Falorsi[1], Giulia Ponzini[2], Paolo Righi[3]

1. INTRODUCTION

Longitudinal studies, based on repeated observations of the same statistical units over time, represent an invaluable source for analysing the current state and the changes in human populations over time.

Longitudinal studies traditionally comprehend panel studies, planned according to specific periodicity and time length, cohort studies based on people with shared experience (e.g. master's degree or first maternity) or characteristics at a particular time point (year of birth), retrospective studies based on different sources regarding past times. Victorian Britain used panel opinions to make better decisions in the nineteenth century (xxx). In the 1950s, we saw a lot of progress with panel studies, which were used to track client satisfaction with enterprises. Panel studies include a wide range of topics, including health, psychology, sociology, education, income, housing, and work experiences. In the 19th century, Victorian Britain collected panel opinions for better decisions. In 1950th, we had a significant development with panel studies for monitoring the customers' satisfaction with businesses. The main fields of panel studies vary from health, psychology, sociology and education to income, housing and job experiences. Also, the relevant price index survey can be considered a panel survey where the observational units are goods and services over time.

Relevant examples at the EU and North-American levels are the National panel survey (ONS), the European labour force quarterly survey the American SIPPs survey.

We cite the Living Standard Measurement Study (LSMS) (World Bank, 2021). This survey provides information on health, access to essential services (water, etc.), risk of malnutrition, poverty status, etc., for over 50 developing countries. The LSMS Integrated Surveys on Agriculture (LSMS-ISA) is another example of a panel-based survey in several Sub-Saharan African countries which disseminates household panel data with a strong focus on agriculture.

In this work, we focus on the case of panels with a rotating sample design. This case represents a powerful hybrid solution for facing the sample erosion for deaths and movers and the impact of lack of sample representativeness for new births, migration flows. Moreover, the sample fatigue introduces an increasing measurement error. Strengths and weaknesses of panel surveys as the comparisons between longitudinal and cross-sectional studies have been well deepened in the theoretical literature on observational studies.

As the length of the panel surveys increases, there is an increasing interest, but also increasing challenges in preserving the quality of the panel sample estimates. In panel surveys, the accuracy of the estimates depends on several factors common to survey sampling. The effect is particularly evident in long run panels.

A correct design, implementation, and use of a panel survey shall consider a set of methods to deal with these problems at different stages of the statistical process: The sampling design, the data collection, and the estimation.

[1] Piero.falorsi@gmail.com. International consultant.
[2] gponzini@worldbank.org. World Bank.
[3] parighi@istat.it. Istat.

The *sampling design* shall minimize the negative impact on the quality of decreasing representativeness of panel data due to new entrants in the population or missing data due to panel attrition and movers, as well as partially overcome the bias introduced by sample fatigue. Practical approaches to this are the adoption of refreshed samples (i.e. rotating panel or split panels) and the definition of criteria for the inclusion of new individuals in the survey to capture some of the population dynamics.

Movers represent a dynamic sub-population, relevant to capture for describing the changes but complex and costly to be interviewed. Based on tracking rules to follow the movers, the *Data collection* shall retrieve some essential information on individuals or households who dropped out from the survey observation. These actions aim to identify linking variables and other auxiliary variables to be collected for computing the direct sampling weights. Moreover, it is helpful to gather a minimal set of variables on the movers (by a proxy interview) and the non-respondents (by doorstep interview). Finally, a multi-mode data collection (i.e., face to face interview and telephone interview for not tracked movers) could enhance the quality of the survey estimates.

The *Estimation* shall consider the drop-outs, movers, new individuals, and the target population's dynamic over time. This result is achieved through (i) the definition of a weighting process to up-date the direct sampling weights (obtained as the inverse of the inclusion probabilities) by the inclusion of new individuals in the panel-households (Generalized Weight Share Method; Lavalleé, 2007), and (ii) the use of calibration estimators (Singh and Mohl, 2001) with up-to-date known population totals.

The present work addresses the problem of the quality of panel surveys considering the three aspects just illustrated. We consider the estimation at the current time of cross-sectional and longitudinal parameters of a target population. The available data are obtained from the follow-up of statistical units surveyed on previous surveys. In particular, we considered two surveys. Although simplified, this setting clarifies many aspects of the representativeness of the data collected from panel surveys for the current-time estimate of cross-sectional data and flow data. It applies very well to the cases that can characterize the current large-scale panel surveys on households. However, we can easily extend it to more complex issues in actual survey settings. In order to be practical, we consider the specific case of the LSMS-ISA data, using the 2009, 2013 and 2015 waves of the Uganda National Panel Survey as case-study.

The paper is structured as follows: **section 2** introduces panel surveys, defining the purposes and the most critical aspects concerning the quality of the estimates. S**ection 3** gives a formal definition of longitudinal and cross-sectional populations and the target parameters to be estimated in longitudinal studies. **Section 4** illustrates the sampling framework and the estimator based on multisource (Mecatti, …) and indirect sampling (Lavallé, Falorsi, …). It allows dealing with the different representativeness of various sampling sub-populations and the changes in household's composition over time**. Section 5** presents some field operations and the data collection improvements to facilitate the implementation of the proposed methodology. **Section 6** illustrates the empirical application to the Uganda Data. **Section 7** summarizes the main findings and concludes.

2. FORMAL DEFINITION OF CROS-SECTIONAL AND LONGITUDINAL POPULATIONS

2.1. Populations observed at different time points

Let U_t be the population of N_t individuals at the current time t of a given country. U_t is partitioned into M_t subpopulations, of households denoted as $U_{t,1}, \dots, U_{t,i}, \dots, U_{t,M}$. We

represent the set of households as $H_t = \{1, \dots, i, \dots, M_t\}$. The household $U_{t,i}$ has $N_{t,i}$ individuals, being $N_t = \sum_{i=1}^{M_t} N_{t,i}$.

We may denote U_t and H_t as ***cross-sectional populations*** since they refer to a specific point in time.

For analysing the dynamics of cross-sectional populations in a longitudinal setting from an initial point in time denoted as t^*, to the current time t, we have to introduce the concept of ***longitudinal-populations***. For making it more straightforward, let us introduce first this notion referring to individuals, and let $U_{t\leftarrow t^*}$ denote the longitudinal population of $N_{t\leftarrow t^*}$ individuals. The longitudinal-population identification is complicated depending on both the scope of study and longitudinal observation's operability over time (Helliot *et al.* 2009). The most common definitions are the *intersection-population*, $U_{t\leftarrow t^*} = U_{t^*} \cap U_t$, and the *union-population* $U_{t\leftarrow t^*} = U_{t^*} \cup U_t$.

The intersection-population is a proper subset of both U_{t^*} and U_t, and includes the individuals of U_{t^*} who are still living and resident in the country at the time t. The longitudinal survey becomes simpler because we observe the same individuals over time, excluding both births and deaths over the study life. However, we suffer some disadvantages related to the representativity in the current time and to the study of the changes in households' compositions.

The union-population seems best apt to analyse a genuinely dynamic population but obviously requires a correspondingly dynamic sample design to introduce births and entrants so avoiding bias.

In this study, we propose a population definition that is a compromise between the two approaches, defined above, but adopts the advantages from both of them. With this definition, the $N_{t\leftarrow t^*}$ individuals of $U_{t\leftarrow t^*}$, include:

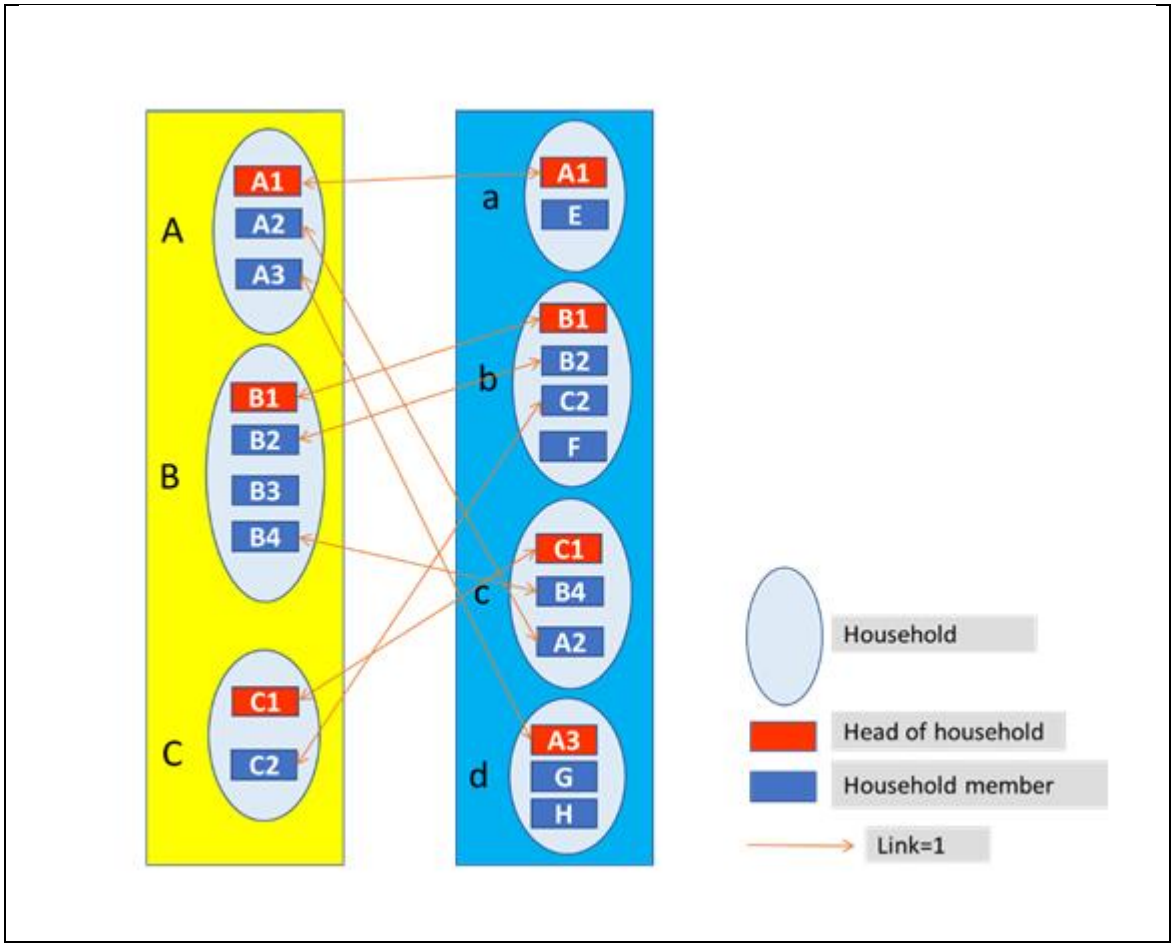
- all the people of *the intersection-population* $U_{t^*} \cap U_t$;
- in addition to the intersection population, $U_{t\leftarrow t^*}$ comprises all new members of their households at time t, even if they were not part of U_{t^*}.

This definition allows considering part of the new entries into the population from time t^*; at least the new-borns and that part of the immigrants who at time t turn out to be members of the households of the individuals of the intersection-population. If immigration to the country is a numerically insignificant phenomenon, then the longitudinal population $U_{t\leftarrow t^*}$ well approximates the cross-sectional population at the current time U_t. Moreover, the longitudinal observation of $U_{t\leftarrow t^*}$ is conceptually simple. We select a sample at the initial time, t^*, and then in the following survey's occasions, we observe all the components of the households of the initially sample-selected people.

Given the definition of the longitudinal population of the individuals, we then may define the longitudinal population of households, denoted below as $H_{t\leftarrow t^*}$, with $M_{t\leftarrow t^*}$ families. We have to determine first the concept of longitudinal household.

The households over time may experience three types of change: disappearance, fusion or division. These changes directly affect cross-sectional and longitudinal analysis and can affect the sample's representativeness to a significant extent (FAO, 2015). Consider the three households A, B and C at the time 1 illustrated in Figure 3.1 below. At the time 2, we find the family A members in three different households: A2 joined household c, A3 formed a new individual household d, and A1 joined household with new member E. Furthermore, the member B3 of household B has disappeared at the time 2.

Figure 3.1. Longitudinal households over time (yellow referred to t^*, blue referred to t)



The picture above well illustrates the problem of defining the longitudinal household over time. We may adopt two broad approaches.

The traditional method is the one-to-one. One household of the time t^* generates only one household at the time t. The continuity rules for identifying the longitudinal family at the time t may differ (Falorsi et al., 2009). With this approach, we risk excluding from the longitudinal analysis some of the households in which the people of $U_{t\leftarrow t^*}$ live. In figure 3.1, we see that if the continuity rule identifies as the longitudinal household the one with the same head of the time t^*, we exclude from the analysis the household d.

To overcome these problems, we propose to take the one-to-many approach. One household of the time t^* generates many families at the time t. Conversely, a family at the time t, may derive from many households at the time t^*. The households of the population $H_{t\leftarrow t^*}$ are those containing people of $U_{t\leftarrow t^*}$. In figure 3.1, we see that the household A of time 1 continues at time 2 with the households a, b and d. With the one-to-many criterion, the household k of U_t is one of the longitudinal households which derive from the household ℓ at the time t^*, iff

$$(3.1) \quad \sum_{j=1}^{N_{t^*,\ell}} \sum_{i=1}^{N_{t,k}} l_{ki,\ell j}^{t\leftarrow t^*} = L_{k,\ell}^{t\leftarrow t^*} > 0,$$

where $l_{ki,\ell j}^{t\leftarrow t^*}$ is the $link\{0,1\}$ variable which assumes value 1 if the individual j of the household ℓ at the time t^* is the same individual i of the household k of the time t. With this approach, there is a perfect correspondence between the longitudinal population of people and that of households. For each person of $U_{t\leftarrow t^*}$, we may define their household at the initial time, t^*, and

that at the current time t. Definition (3.1) includes as a particular case the one-to-one continuity rule. In this case, $L_{k,\ell}^{t\leftarrow t^*}$ is a {0,1} dichotomous variable, which equals one if the household k is the only one household of U_t which is the longitudinal continuation of the household ℓ of U_{t^*}.

2.2. Parameters of interest

2.2.1. Cross-sectional parameters

Let $\mathcal{y}$ and $\mathcal{Y}$ be respectively two quantitative variables observed on individuals and households. The parameter of interest for the inference is the total Y_t referred to the population U_t where:

$$(3.2)\quad Y_t = \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} y_{t,ki} = \sum_{k=1}^{M_t} Y_{t,k}$$

in which $y_{t,ki}$ is the value of $\mathcal{y}$ of the individual i of the household k, and

$$(3.3)\quad Y_{t,k} = \sum_{i=1}^{N_{t,k}} y_{t,ki}$$

is the total of $\mathcal{Y}$ for the household k.

Now we assume that the $\mathcal{y}$ variable is a categorical variable with value $y_{t,ki} = p$ $(p = 1, \dots, P)$ if the person i of the household k belong to the $p-th$ category. For instance, if the variable $\mathcal{y}$ describes the employment status with three categories (employed, unemployed, not labour force), $y_{t,ki} = 1$ denotes the status of being employed. Let $\mathcal{y}^p$ $(p = 1, \dots, P)$ denote a dichotomous (0,1) variable which, for the individual i of the household k, is equal to the value $y_{t,ki}$, where

$$y_{t,ki}^p = \begin{cases} 1 & \text{if } y_{t,ki} = p \\ 0, & \text{otherwise} \end{cases}.$$

The parameters of interest are the frequency totals of the different P categories of the variable $\mathcal{y}$:

$$(3.4)\quad Y_t^p = \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} y_{t,ki}^p = \sum_{k=1}^{M_t} Y_{t,k}^p \qquad (p = 1, \dots, P),$$

where $Y_{t,k}^p$ indicates the total of the variable $\mathcal{y}^p$, that is the number of individuals in the household k characterized by the category p of the variate $\mathcal{y}$.

Finally, we can be interested in variables typically defined at the household level describing a specific household condition (poverty or non-poverty, the status of malnutrition, etc.). We indicate with $\mathcal{Y}$ the categorical variable with values $Y_{t,k} = p$ $(p = 1, \dots, P)$ if the household $k \in U_t$ belongs to the p-th category and we denote as $\mathcal{Y}^p$ $(p = 1, \dots, P)$ the dichotomous (0,1) variables with values

$$Y_{t,k}^p = \begin{cases} 1 & \text{if } Y_{t,k} = p \\ 0, & \text{otherwise} \end{cases}.$$

In this case the parameter in (3.4) counts the number of households in the $p-th$ condition.

2.2.2. Longitudinal parameters of individuals

The **net change** parameter is given by

$$(3.5)\quad \Delta_{t\leftarrow t'}^p = Y_t^p - Y_{t^*}^p .$$

Parameter (3.5) expresses the difference between the cross-sectional total at the time t with the corresponding total at the time t'. Analysis of net change using cross-sectional aggregate estimates may hide important gross changes occurring at the individual level, which may be revealed from longitudinal data. (Steel et al. 2009). A longitudinal survey, following the same people through repeated interviews, can be used to estimate yearly trends as well as persistence (Lohor, 2009). The net change reflects changes in both the characteristics and composition of the population.

The **gross-change** implies the measurement of the phenomenon on the same units; thus, it must consider unaltered units in their definition over time.

The gross-change of individuals who in the time t^* were in the category g of the variable y, and are in the category p of the same variable (with $g, p = 1, \dots, P$) in the current time t can be expressed as:

$$(3.6) \quad \overleftarrow{Y}^{p,g}_{t\leftarrow t^*} = \sum_{\ell=1}^{M_{t'}} \sum_{j=1}^{N_{t^*,\ell}} \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} l^{t\leftarrow t^*}_{ki,\ell j}\, y^{p}_{t,ki}\, y^{g}_{t^*,\ell j} = \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} \overleftarrow{y}^{p,g}_{t\leftarrow t^*,ki},$$

where

$$(3.7) \quad \overleftarrow{y}^{p,g}_{t\leftarrow t^*,ki} = \sum_{\ell=1}^{M_{t'}} \sum_{j=1}^{N_{t^*,\ell}} l^{t\leftarrow t^*}_{ki,\ell j}\, y^{p}_{t,ki}\, y^{g}_{t^*,\ell j}$$

is the flow variable which equals 1 if the individual ki changed from p to g in time interval $(t' - t)$ and 0 otherwise. Thus, the gross-change may be expressed as a cross-sectional total (at the time t) of the *flow-variable* $\overleftarrow{y}^{p,g}_{t\leftarrow t^*,ki}$. This variable can be measured directly only on the units of the intersection population, $U_t \cap U_{t^*}$.

3.2.3. *Changes of the households over time*

The **net-change** of households may be expressed as already given in (3.5) except the fact that the totals Y^p_t and $Y^p_{t'}$ are referred to the variable $\mathcal{Y}^p$ above introduced.

The **gross-change** measures the amount of change conditions: how many households shift from negative to positive condition (poverty to non-poverty, malnutrition to non- malnutrition) or from positive to negative condition (non-poverty to poverty, non-malnutrition to malnutrition) or eventually remain in the same positive or negative condition. As for the individuals, the gross-change is observed on the categorical variables.

The gross-change of households that changed from the category g to the category p of the variate $\mathcal{Y}$ (with $g, p = 1, \dots, P$) can be expressed as

$$(3.8) \quad \overleftarrow{Y}^{p,g}_{t\leftarrow t'} = \sum_{k=1}^{M_t} \sum_{\ell=1}^{M_{t^*}} \frac{L^{t\leftarrow t^*}_{k,\ell}}{L^{t\leftarrow t^*}_{k}} Y^{g}_{t^*,\ell} Y^{p}_{t,k} = \sum_{k=1}^{M_t} \overleftarrow{Y}^{p,g}_{t\leftarrow t^*,k},$$

in which

$$(3.9) \quad L^{t\leftarrow t^*}_{k} = \sum_{\ell=1}^{M_{t^*}} \sum_{j=1}^{N_{t^*,\ell}} \sum_{i=1}^{N_{t,k}} l^{t\leftarrow t^*}_{ki,\ell j}$$

indicate the number of ***potential links*** of the household k. This quantity denotes the number of individuals of the household k which belong to U_{t^*}. It plays a central role for both the definition of the longitudinal parameter of the *gross-change* and for producing unbiased estimates of the parameters of interest (see section 3). In order to clarify how to calculate this number in the

following Schema 3.1, we provide the value of $L_k^{t\leftarrow t^*}$ for each of the four households (**a**, **b**, **c** and **d**) of the time 2 given in picture 3.1.

Schema 3.1. Values of $L_k^{t\leftarrow t^*}$ for the households at the time 2 in Picture 2.1

Household	$L_k^{t\leftarrow t^*}$	
	One to one	One to many
A	1	1
B	1	3
C	1	3
d	0	1

If $L_k^{t\leftarrow t^*} = L_{k,\ell}^{t\leftarrow t^*}$, then the household k is the only one household of the time t which derives from the original household ℓ of U_{t^*}.

Reconsidering now Formula (3.8), we note that it can be expressed as

$$(3.10)\quad \overleftarrow{Y}_{t\leftarrow t^*}^{p,g} = \sum_{k=1}^{M_t} \overleftarrow{Y}_{t\leftarrow t^*,k}^{p,g},$$

where

$$(3.11)\quad \overleftarrow{Y}_{t\leftarrow t^*,k}^{p,g} = \sum_{\ell=1}^{M_{t^*}} \frac{L_{k,\ell}^{t\leftarrow t^*}}{L_k^{t\leftarrow t^*}} Y_{t^*,\ell}^{\ g} Y_{t,k}^{p},$$

is the *flow-variable* indicating the change in the time interval $(t^* - t)$ of the *household k* (as identified at the time t) from the category g to the category p of the variable $\mathcal{Y}$.

In the case of the one-to-many continuity rule, $\overleftarrow{Y}_{t\leftarrow t^*,k}^{p,g}$ is a real value variable defined in the [0,1] interval.

In the case of one-to-one continuity rule, $\overleftarrow{Y}_{t\leftarrow t^*,k}^{p,g}$ is a $\{0; 1\}$ dichotomous variable.

The schema 3.2. shows the computation of the household gross-change according to the continuity rules and the household relationships depicted in the figure 3.1.

Schema 3.2. Gross-change parameter computation according to the household continuity rule following figure 2.1

One to one continuity rules*

Household	t^*	a	b	c	D
t	Poverty status at given time^	0	0	1	0
A	0	1	0	0	0
B	1	0	1	0	0
C	1	0	0	1	0

Household $U_{2\leftarrow 1}$	Computation	Parameter	%
Positive change event	1 (Bb)	1	33.3
Negative change event	0	0	0.0
Stationary event	1 (Aa)+ 1(Cc)	2	66.7
Total households		3	100

One to many continuity rule

Household	t^*	a	b	c	d
t	Poverty status at given time^	0	0	1	0
A	0	1/3	0	1/3	1
B	1	0	2/3	1/3	0
C	1	0	1/3	1/3	0

Household $U_{2\leftarrow 1}$	Computation	Parameter	%
Positive change event	2/3 (Bb)+1/3(Cb)	1	25.0
Negative change event	1/3 (Ac)	1/3	8.3
Stationary event	1 (Aa)+1(Ad)+1/3 (Bc)+ 1/3(Cc)	8/3	66.7
Total households		4	100.0

*Continuity rule follow the head of household
^Cross-sectional poverty status=1, non-poverty status=0

^Cross-sectional poverty status=1, non-poverty status=0

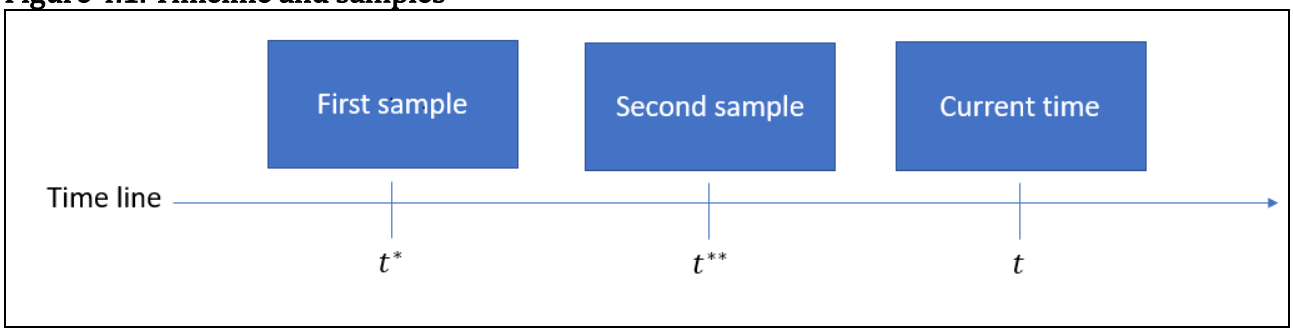
3. SAMPLING SETTING AND ESTIMATOR

3.1. Sampling design

For illustrating the methodology we consider three specific points in the time

- ✓ **Initial time** t^*. At this time, we start the longitudinal survey, by selecting the **first** sample of households.
- ✓ **Intermediate time** t^{**}. At this time, we select the **second** sample of households and make the follow-up at the time t of the first sample.
- ✓ **Current time** t. At this follow-up at the time t of the first sample and second sample.

Figure 4.1. Timeline and samples



We summarize here below the main sampling operations carried out at the different time points.

First time $\boldsymbol{t^*}$. We select a fixed sample size of households, S_{t^*}, from the cross-sectional population H_{t^*}, where the household $\ell \in H_{t^*}$ enters in the sample with the inclusion probability $\pi_{t^*,\ell}$. S_{t^*} includes m_{t^*} households, being $m_{t^*} = \sum_{\ell \in H_{t^*}} \pi_{t^*,\ell}$.

We observe in the sample all the $N_{t^*,\ell}$ individuals of the sample household ℓ. Thus, the $j-$ th $(j = 1, \dots, N_{t^*,\ell})$ member of the household ℓ is included in the sample with the same probability of inclusion of their household. Let

$$n_{t^*} = \sum_{\ell=1}^{m_{t^*}} N_{t^*,\ell}$$

denote the realized sample size of S_{t^*} in terms of individuals. While m_{t^*} is fixed for each sampling selection, n_{t^*} is a random outcome depending on the number of people in the sample families.

Second time $\boldsymbol{t^{**}}$. We select a new sample, $S_{t^{**}}$, of households from the cross-sectional population $H_{t^{**}}$. $S_{t^{**}}$ is selected independently from S_{t^*} and is a fixed-size sample of $m_{t^{**}}$ households. The household $k \in H_{t^{**}}$ enters in the sample with the inclusion probability $\pi_{t^{**},k}$. We observe in the sample all the $N_{t^{**},k}$ individuals of the sample household k. In this way we observe a sample of $n_{t^{**}}$ of individuals.

Moreover, the enumerators track all the n_{t^*} individuals of the sample S_{t^*}, according to the survey-tracking rules. For instance, in the Uganda LSMS, all the people who have not moved from the original Parish are tracked. Once the individual selected in the original panel is re-contacted for the survey, the enumerator identifies all the family members and collects the survey data on them. This way ensures the study considers the changes in the households'

composition (due to births, deaths, marriages, etc.) of the n_{t^*} people involved in the survey at the previous time t^*. Thus, the enumerators form the *longitudinal sample* $S_{t^{**}\leftarrow t^*}$ of $m_{t^{**}\leftarrow t^*}$ households with $n_{t^{**}\leftarrow t^*}$ individuals.

Current time *t*. The enumerators track all the $n_{t^{**}}$ individuals of the sample $S_{t^{**}}$, according to the survey-tracking rules. In this way, the enumerators form the *longitudinal sample* $S_{t\leftarrow t^{**}}$. This sample has $m_{t\leftarrow t^{**}}$ households and $n_{t\leftarrow t^{**}}$ people.

Furthermore, the enumerators track all the people of $S_{t^*} \cap S_{t^{**}\leftarrow t^*}$ who are the people selected in the original sample S_{t^*} and yet observed in the longitudinal sample $S_{t^{**}\leftarrow t^*}$. So, the enumerators form the *longitudinal sample* $S_{t\leftarrow t^*}$. This sample has $m_{t\leftarrow t^*}$ households and $n_{t\leftarrow t^*}$ people.

The samples for the estimation at the time t are $S_{t\leftarrow t^{**}}$ and $S_{t\leftarrow t^*}$.

3.2. Movers and representativeness of the samples

Movers create two types of problems in longitudinal surveys. First, there is a need to collect new contact information and the associated increased risk of failure to contact the sample unit. Second, there are additional costs. LSMS surveys that employ face-to-face interviews use cluster and multi-stage sampling designs to control costs, so following a mover to a new address may incur considerable extra cost if the new address is not in one of the original sample areas. Also, there is a risk that no interviewer may be available to visit a mover if the move is discovered only during the field operations. Hence, a tracking protocol needs to be put in place to **track** and **record** movers among the sample members. Thomas *et al.* (2001) highlight some surveys in developing countries that suffered from substantial attrition due to failure to track movers. However, omitting movers may create an obvious bias in surveys.

The target current population of households can be defined as the union of the three disjoint subpopulations:

(4.1) $\quad H_t = F_{t\leftarrow t^*} \cup \mathcal{E}_{t^{**}\leftarrow t^*} \cup \mathcal{E}_{t\leftarrow t^{**}}$

where

- ✓ $F_{t\leftarrow t^*}$ is the subpopulation of the households which remain fixed from the time t^* or being traceable from the enumerators from time t^* to time t. For instance, according to the tracking rules adopted for the LSMS in Uganda, the traceable households are those in which at least one individual was already present at the time t^* and has remained in the original Parish till time t. $F_{t\leftarrow t^*}$ does not include the households of all immigrants, which contain only the people who entered the country after the time t^*, and the families consisting only of people who moved after the time t^* and have not been traceable.
- ✓ $\mathcal{E}_{t^{**}\leftarrow t^*}$ comprises the **households** which members consist only of **immigrants** or **movers** in the time interval $[t^*, t^{**}]$.
- ✓ $\mathcal{E}_{t\leftarrow t^{**}}$ includes the **households** the members of which are only **immigrants** or **movers** in the time interval $[t^{**}, t]$.

Partition (4.1) makes it necessary to define the **representativeness** of the two longitudinal samples ($S_{t\leftarrow t^{**}}$ and $S_{t\leftarrow t^*}$) for as regards di sub-populations of H_t. With the term representativeness, we intend that the sample data allow building direct unbiased estimates of a given population. We have the situation illustrated in the figure 4.2.

For the **Sample $S_{t\leftarrow t^*}$**, the tracking rules limit the representativeness of the sample $S_{t\leftarrow t^*}$ only to the sub-population

$$F_{t\leftarrow t^*}$$

of households of the time t who have been traceable from time t^* to time t.

For the **Sample $S_{t\leftarrow t^{**}}$**, the tracking rules restrain the representativeness of the sample data only to the sub-population

$$F_{t\leftarrow t^{**}}$$

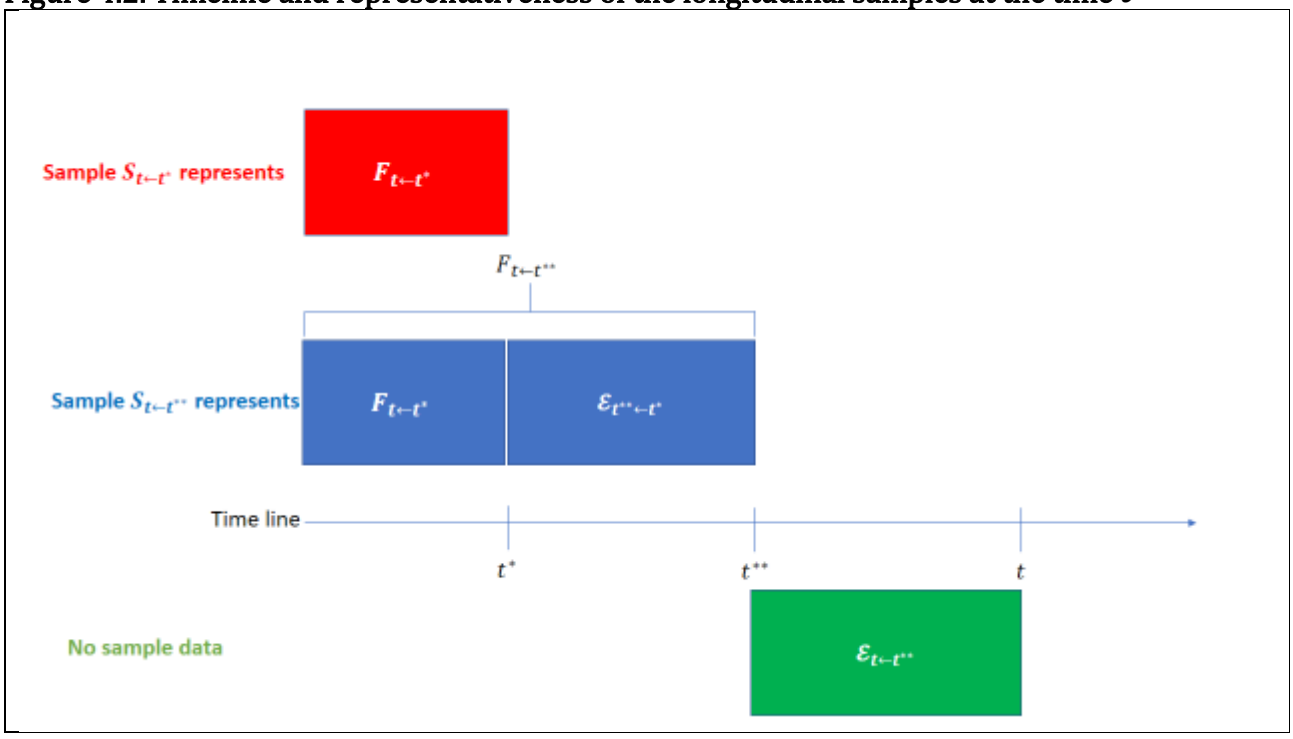
of households of the time t who have been traceable from the time t^{**} to the time t.

$F_{t\leftarrow t^{**}}$ can be partitioned into the two subpopulations $F_{t\leftarrow t^*}$ and $\mathcal{E}_{t^{**}\leftarrow t^*}$, being

$$F_{t\leftarrow t^{**}} = F_{t\leftarrow t^*} \cup \mathcal{E}_{t^{**}\leftarrow t^*}.$$

From the above, we see that the sample data do not allow to compute direct estimates of the subpopulation $\mathcal{E}_{t\leftarrow t^{**}}$.

Figure 4.2. Timeline and representativeness of the longitudinal samples at the time t



Y_t may be expressed as the sum of three addenda, each of which is the total of y related to one of the three sub-populations $F_{t\leftarrow t^*}$, $F_{t\leftarrow t^{**}}$, and $\mathcal{E}_{t\leftarrow t^{**}}$

$$(4.2)\quad Y_t = Y_{F_{t\leftarrow t^*}} + Y_{\mathcal{E}_{t^{**}\leftarrow t^*}} + Y_{\mathcal{E}_{t\leftarrow t^{**}}}$$

being

$$(4.3a)\quad Y_{F_{t\leftarrow t^*}} = \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} y_{t,ki}\, F_{k,t\leftarrow t^*} = \sum_{k=1}^{M_t} Y_{t,k} F_{k,t\leftarrow t^*},$$

$$(4.3b)\quad Y_{\mathcal{E}_{t^{**}\leftarrow t^*}} = \sum_{k=1}^{M_t} \sum_{i=1}^{N_{t,k}} y_{t,ki}\, \mathcal{E}_{k,t^*\leftarrow t^*} = \sum_{k=1}^{M_t} Y_{t,k} \mathcal{E}_{k,t^*\leftarrow t^*},$$

$$(4.3c)\quad Y_{\mathcal{E}_{t\leftarrow t^{**}}} = \sum_{k=1}^{M_t}\sum_{i=1}^{N_{t,k}} y_{t,ki}\left[\left(1-F_{k,t\leftarrow t^*}\right)\left(1-\mathcal{E}_{k,t^{**}\leftarrow t^*}\right)\right]$$
$$= \sum_{k=1}^{M_t} Y_{t,k}\left[\left(1-F_{k,t\leftarrow t^*}\right)\left(1-\mathcal{E}_{k,t^{**}\leftarrow t^*}\right)\right],$$

where, $F_{k,t\leftarrow t^*}$ and $\mathcal{E}_{k,t^{**}\leftarrow t^*}$ are two dichotomous variables defined at the household level being

$$(44)\quad F_{k,t\leftarrow t^*} = \begin{cases} 1 & \text{if } k \in F_{t\leftarrow t^*} \\ 0, & \text{otherwise} \end{cases}, \quad \mathcal{E}_{k,t^{**}\leftarrow t^*} = \begin{cases} 1 & \text{if } k \in \mathcal{E}_{t^{**}\leftarrow t^*} \\ 0, & \text{otherwise} \end{cases}.$$

The representativeness of the samples allows to compute a direct unbiased estimate only for the aggregate:

$$(4.5)\quad Y_{F_{t\leftarrow t^{**}}} = Y_{F_{t\leftarrow t^*}} + Y_{\mathcal{E}_{t^{**}\leftarrow t^*}}.$$

$Y_{F_{t\leftarrow t^{**}}}$ represents an accurate measure of the total Y_t only if the aggregate $Y_{\mathcal{E}_{t\leftarrow t^{**}}}$ is negligible. This condition holds if the time point t^{**} is close to the current time t.

3.3. Estimation process

3.3.1. Direct estimator

The direct estimate of the total $Y_{F_{t\leftarrow t^{**}}}$, as expressed by formula (4.5), is given by:

$$(4.6)\quad \hat{Y}_{F_{t\leftarrow t^{**}}} = \alpha\hat{Y}^{S_{t\leftarrow t^*}}_{F_{t\leftarrow t^*}} + (1-\alpha)\hat{Y}^{S_{t\leftarrow t^{**}}}_{F_{t\leftarrow t^*}} + \hat{Y}^{S_{t\leftarrow t^{**}}}_{\mathcal{E}_{t^{**}\leftarrow t^*}},$$

in which $\hat{Y}^{S_{t\leftarrow t^*}}_{F_{t\leftarrow t^*}}$, is the generalized weight share method (GWSM, Lavallé, 2007) estimator of the total $Y_{F_{t\leftarrow t^*}}$ obtained with the data collected by the longitudinal sample $S_{t\leftarrow t^*}$, $\hat{Y}^{S_{t\leftarrow t^{**}}}_{F_{t\leftarrow t^*}}$, and $\hat{Y}^{S_{t\leftarrow t^{**}}}_{\mathcal{E}_{t^{**}\leftarrow t^*}}$ are the GWSM estimates of the totals $Y_{F_{t\leftarrow t^*}}$ and $Y_{\mathcal{E}_{t^{**}\leftarrow t^*}}$ computed with the data of the sample $S_{t\leftarrow t^{**}}$, with $0 \le \alpha \le 1$. The parameter α can either be fixed in advance or calculated from the survey data. Further discussion on the choice of α is provided below in formula (4.15).

The GWSM estimator $\hat{Y}^{S_{t\leftarrow t^*}}_{F_{t\leftarrow t^*}}$ is given by:

$$(4.7)\quad \hat{Y}^{S_{t\leftarrow t^*}}_{F_{t\leftarrow t^*}} = \sum_{k=1}^{m_{t\leftarrow t^*}}\sum_{i=1}^{N_{t,k}} y_{t,ki}\, d_{S_{t\leftarrow t^*},k}$$

where $d_{S_{t\leftarrow t^*},k}$ is the GWSM weight, given by

$$(4.8)\quad d_{S_{t\leftarrow t^*},k} = \frac{1}{L_k^{t\leftarrow t^*}}\sum_{\ell\in S_{t^*}}\sum_{i=1}^{N_{t^*,\ell}}\frac{1}{\pi_{t^*,\ell}} l^{t\leftarrow t^*}_{ki,\ell j}.$$

The enumerators can collect the value of $L_k^{t\leftarrow t^*}$ during the survey, gathering the information on how many members of the family were present in the country at the time t^*. Furthermore, we note that all the households in sample $S_{t\leftarrow t^*}$ have a value of $F_{k,t\leftarrow t^*}$, which equals 1.

The GWSM estimators $\hat{Y}^{S_{t\leftarrow t^{**}}}_{F_{t\leftarrow t^*}}$ and $\hat{Y}^{S_{t\leftarrow t^{**}}}_{\mathcal{E}_{t^{**}\leftarrow t^*}}$ are given by:

$$(4.9)\quad \hat{Y}^{S_{t\leftarrow t^{**}}}_{F_{t\leftarrow t^*}} = \sum_{k=1}^{m_{t\leftarrow t^{**}}}\sum_{i=1}^{N_{t,k}} y_{t,ki}\, F_{k,t\leftarrow t^*} d_{S_{t\leftarrow t^{**}},k},$$

(4.10) $\hat{Y}_{\mathcal{E}_{t^{**}\leftarrow t^*}}^{S_{t\leftarrow t^{**}}} = \sum_{k=1}^{m_{t\leftarrow t^{**}}} \sum_{i=1}^{N_{t,k}} y_{t,ki} \left(1 - F_{k,t\leftarrow t^*}\right) d_{S_{t\leftarrow t^{**}},k}$,

where $d_{S_{t\leftarrow t^{**}},k}$ is the GWSM weight, given by

(4.11) $d_{S_{t\leftarrow t^{**}},k} = \frac{1}{L_k^{t\leftarrow t^{**}}} \sum_{\ell \in S_{t^{**}}} \sum_{i=1}^{N_{t^{**},\ell}} \frac{1}{\pi_{t^{**},\ell}} l_{ki,\ell j}^{t\leftarrow t^{**}}$.

The enumerators can collect the value of $L_k^{t\leftarrow t^{**}}$ during the survey, gathering the information on how many members of the family were present in the country at the time t^{**}.

For the households collected in the sample $S_{t\leftarrow t^{**}}$, the value of $F_{k,t\leftarrow t^*}$ may be either equal to 0 or 1. If $F_{k,t\leftarrow t^*} = 0$, then $\mathcal{E}_{k,t^{**}\leftarrow t^*} = 1$. The enumerators can collect the value of $F_{k,t\leftarrow t^*}$ during the survey, asking all the people in the family k if:

1. they immigrated to the country after the time t^*,
2. or they moved outside the perimeter established for tracking the sample people, after the time t^*.

If all the members of the household k give a positive answer to question 1, or question 2, then the value of $F_{k,t\leftarrow t^*}$ equals 0.

Let

(4.12) $S_{(t\leftarrow)} = S_{t\leftarrow t^*} \cup S_{t\leftarrow t^{**}}$

be the longitudinal *sample-union* at the current time t, obtained by joining together the two longitudinal samples $S_{t\leftarrow t^*}$ and $S_{t\leftarrow t^{**}}$. The sample $S_{(t\leftarrow)}$ has $m_{(t\leftarrow)} = m_{t\leftarrow t^*} + m_{t\leftarrow t^{**}}$ households, and $n_{(t\leftarrow)} = n_{t\leftarrow t^*} + n_{t\leftarrow t^{**}}$ individuals.

We may express the estimator $\hat{Y}_{F_{t\leftarrow t^{**}}}$ in the standard form as a weighted sum of the data in the sample union as:

(4.13) $\hat{Y}_{F_{t\leftarrow t^{**}}} = \sum_{k \in S_{(t\leftarrow)}} \sum_{i=1}^{N_{t,k}} y_{t,ki}\ d_{S_{(t\leftarrow)},k}$,

where $d_{S_{(t\leftarrow)},k}$ is the direct sample weight of the household k of the sample $S_{(t\leftarrow)}$. With straightforward algebraical manipulation, starting from the above formula (4.6), … , (4.11) we have

(4.14) $d_{S_{(t\leftarrow)},k} = \begin{cases} \alpha d_{S_{t\leftarrow t^*},k} & \text{if } k \in S_{t\leftarrow t^*} \\ (1-\alpha) d_{S_{t\leftarrow t^{**}},k} & \text{if } k \in S_{t\leftarrow t^{**}} \wedge F_{k,t\leftarrow t^*} = 1 \\ d_{S_{t\leftarrow t^{**}},k} & \text{if } k \in S_{t\leftarrow t^{**}} \wedge F_{k,t\leftarrow t^*} = 0 \end{cases}$,

where the expressions for $d_{S_{t\leftarrow t^*},k}$ and $d_{S_{t\leftarrow t^{**}},k}$ are given in formula (4.8) and (4.11) respectively. In the following we denote $d_{S_{(t\leftarrow)},k}$ as the *base weights* of the sample-union.

As of the definition of the α value, Singh and Mecatti (2011) provided an in-depth illustration of the different approaches proposed in literature for fixing the optimal value of α in the context of multiple frame surveys. Hartley (1962, 1974) proposed choosing α to minimize the variance of $\hat{Y}_{F_{t\leftarrow t^{**}}}$. Unfortunately, the solution depends on the variable y and may be negative. Hartley (1974) suggested opting for a simpler alternative expression which is always positive, even if

it depends from y. We suggest here an even simpler solution which does not suffer from the above drawback and well approximates the Hartley's solution

$$(4.15)\ \alpha = \frac{m_{t\leftarrow t^{*}}}{m_{t\leftarrow t^{*}} + m_{t\leftarrow t^{**}}}.$$

If $m_{t\leftarrow t^{*}} \cong m_{t\leftarrow t^{**}}$, as in the case illustrated in Section 5 below, then $\alpha = 0.5$. The reciprocal of α is known as *factor of multiplicity*. In the case illustrated in Section 5 below, this factor equals 2.

3.3.2. Calibration estimator

It could be necessary to amend the direct estimates $\hat{Y}_{F_{t\leftarrow t^{**}}}$ (see expression 3.14) obtained with the direct final weights $d_{S_{(t\leftarrow)},k}$ for adjusting for at least three phenomena which may cause the most relevant loss of accuracy in the estimates: (i) the first phenomenon is the *under-coverage* since $\hat{Y}_{F_{t\leftarrow t^{**}}}$ does not represent the aggregate $Y_{\mathcal{E}_{t\leftarrow t^{**}}}$. (ii) The second problem is the *non-response*. (iii) Moreover, it could be necessary to modify the direct estimators because there are too-large differences in the sample estimates with the known totals of the cross-sectional population H_t. We can pursue the three goals jointly with the *calibration estimator* (Deville and Särndal, 1992; Singh and Mohl, 1996). This estimator defines *calibrated weights,* $w_{S_{(t\leftarrow)},k}$, to be used for the estimation which are the closest as possible to the $d_{S_{(t\leftarrow)},k}$ direct weights and allow to reproduce the know totals of some auxiliary variables. The calibration estimator of the total Y_t is then defined as:

$$(4.16)\quad \hat{Y}_{t,cal} = \sum\nolimits_{k\in S_{(t\leftarrow)}} \sum\nolimits_{i=1}^{N_{t,k}} y_{t,ki}\ w_{S_{(t\leftarrow)},k},$$

or if the variable is referred to the households as

$$(4.16b)\quad \hat{Y}_{t,cal} = \sum\nolimits_{k\in S_{(t\leftarrow)}} Y_{t,k}\, w_{S_{(t\leftarrow)},k}\ .$$

In order to introduce this estimator, let $\mathbf{X}_t$ be a column vector of auxiliary totals known for the population U_t from administrative data or last Census, or demographic statistics. Let $\mathbf{X}_{t,k}$ be a vector of auxiliary variables of the household k such that

$$\sum\nolimits_{k=1}^{M_t} \mathbf{X}_{t,k} = \mathbf{X}_t.$$

Let us suppose that the vector $\mathbf{X}_t$ is known for the whole population and that the vectors $\mathbf{X}_{t,k}$ are known for the sample households.

The calibrated weights $w_{S_{(t\leftarrow)},k}$ are obtained as solution of the following minimum constraint problem

$$(4.17)\quad \begin{cases} \sum_{k\in S_{(t\leftarrow)}} D\left(d_{S_{(t\leftarrow)},k}, w_{S_{(t\leftarrow)},k}\right) = \min \\ \sum_{k\in S_{(t\leftarrow)}} \mathbf{X}_{t,k}\, w_{S_{(t\leftarrow)},k} = \mathbf{X}_t \\ L d_{S_{(t\leftarrow)},k} \le w_{S_{(t\leftarrow)},k} \le U d_{S_{(t\leftarrow)},k} \end{cases},$$

where: $D\left(d_{S_{(t\leftarrow),k}}, w_{S_{(t\leftarrow),k}}\right)$ is a *truncated distance function* between $d_{S_{(t\leftarrow),k}}$, and $w_{S_{(t\leftarrow),k}}$, ; $0 \leq L \leq 1$; and $U \geq 1$. The truncated distance functions (Singh and Mohl, 1996) ensures that the calibrated weights are bounded in the interval $\left(Ld_{S_{(t\leftarrow),k}}, Ud_{S_{(t\leftarrow),k}}\right)$, thus ensuring in finding a solution without outlier or negative weights. The problem (4.17) can be solved, among other solutions, with the open-source software Regenesees[4] which allows the use of two different truncated functions: the truncated linear and the truncated logistic.

The calibration estimator, $\hat{Y}_{t,cal}$, with weights $w_{S_{(t\leftarrow),k}}$, obtained as solution of (4.17), has the following positive characteristics. The system (4.17) defines weights $w_{S_{(t\leftarrow),k}}$ at the household level, and we can use them for estimating parameters of both individuals and households. Thus, they ensure the coherence of the household estimates with those of the individuals. The estimates of the auxiliary variables $\mathbf{X}_{t,k}$ are benchmarked to the known totals (or estimated by accurate a large and accurate survey) $\mathbf{X}_t$, defined at country level. Therefore, the weights $w_{S_{(t\leftarrow),k}}$ guarantee the coherence of the sample estimates with the system of statistics at the country level. Deville and Särndal (*et al.*, 2015) demonstrate the calibration estimators converge in probability to the regression estimator. Therefore, they have a well-known asymptotic behaviour. They produce inferences which are robust and have sound statistical properties concerning both the sampling design and the statistical model. If the auxiliary variables are explicative of the non-response, the calibration estimator reduces the no-response bias (Särndal *et al.*, 2015). The use of truncated distance functions allows the weights $w_{S_{(t\leftarrow),k}}$ are always positive, where the outliers of the weights have a limited impact on the final estimates. The two steps approach (Sarnadal ,..., 2014) improves the accuracy of the estimates when the reasons of the attrition are basically different from the reasons of the sampling list under-coverage. So that we implement two calibrations: the first calibration adjusts the base weights by panel attrition, the second calibration uses the adjusted weights to produces the final weights to tackle the under-coverage.

The calibration estimator of the (3.5) is given by

$$\hat{\Delta}^{p}_{t\leftarrow t',cal} = \hat{Y}^{p}_{t,cal} - \hat{Y}^{p}_{t',cal} .$$

The calibrated estimate of the gross-change for the individuals, $\overleftarrow{Y}^{p,g}_{t\leftarrow t^*}$, as defined by formula (3.6), may be obtained simply as the weighted sum of the flow variables $\overleftarrow{y}^{p,g}_{t\leftarrow t^*,ki}$ defined in expression (3.7).

The calibrated estimate of the gross-change for the households, $\overleftarrow{Y}^{p,g}_{t\leftarrow t^*}$, as defined by formula (3.10) may be obtained simply as the weighted sum of the flow variables $\overleftarrow{Y}^{p,g}_{t\leftarrow t^*,k}$ defined in expression (3.11).

4. DATA COLLECTION

4.1. Tracking rules

The statistical offices define the tracking rules to deal with the sample size reduction due to the movers. The definition of the rules has to consider the trade-off between improving the accuracy of the estimates and the costs to follow the movers in the Country. The trade-off generally leads to a comprise solution that observes the movers within a restricted area. In this case an individual is tracked if they move inside the area of the first observation.

[4] https://www.istat.it /it/metodi-e-strumenti/metodi-e-strumenti-it/elaborazione/strumenti-di-elaborazione/regenesees, accessed September 2019

In order to carry out the tracking, the survey has to define a tracking protocol based on:

- the **kind of movers** to be tracked. In order to describe the procedure in concrete terms, let us consider three waves t^*, t^{**} and t already introduced in Section 4.1. At the beginning of the field operations for the current time t, we have the sample of individuals and households originally selected at t^* and t^{**} or the individuals incorporated in their households in the successive waves. The survey plan applies the tracking protocol at t only to the individuals selected in the original sample t^* and t^{**}.
- The **delimitation of the area** where the tracking is carried out (i.e. Enumeration Area, District, Parish, etc.).
- The **questionnaire definition**, which has to include useful questions for contacting the people for the next waves (e.g. whether the respondent thinks to move in the next wave of the panel; the reason of the moving event; the phone number of the mover).
- The protocol for the **proxy interviews** defining the minimal set of variables to be collected on non-respondents by doorstep interview and on movers by a proxy interview.
- The definition of the **data collection process for movers** which considers the area of interview (e.g. face to face interview if movers stays in the sample area, phone interview if mover go in another area).

4.2. Linking and other auxiliary variables for computing the base weights

We denote as base weight, the weight computed according to the Generalized Weight Share Method (GWSM; Lavalleé, 2007) divided by the multiplicity factor (Section 4.3.1). These weights are defined at household level and each individual has the base weight of the household in which they belong to.

According to Section 4.3.1, the GWSM requires: the variable linking the individual to the original sample in which belongs to. The variable linking the individual to the target population, that means the year of appearing in the target population (for new-borns and immigrants).

The process of computation of the base weights proceeds by taking the multiplicity factor into account. The multiplicity factor represents the number of chances of a household to be selected in the original samples. Each refresh sample is a potential chance of the units to be selected. According to the (4.15) if $m_{t\leftarrow t^*} \cong m_{t\leftarrow t^{**}}$, we assign the multiplicity factor to the household observed at the time t according to the two conditions below:

condition 1. If the family has at least one component that belongs to the target population in the wave t^* and is not a mover (or non-traceable mover), the multiplicity factor is 2.

condition 2. If no component of the household satisfies condition 1) then the multiplicity factor is 1.

The multiplicity factor equals 2 if at least one component was selected in the original sample at the wave t^*. The multiplicity factor equals 2 if at least a component of the household was selected at the wave t^{**}, and variable $F_{k,t\leftarrow t^*}$ of the household equals 1. The multiplicity factor equals 1 if at least a component of the household was selected at the wave t^{**}, and variable $F_{k,t\leftarrow t^*}$ of the household equals 0.

Finally, the response indicator variables of the individuals in the current wave and in the first wave (for the individuals selected in the original samples) complete the information to build the base weights.

4.3. Defining the minimal set of variables to be collected on non-respondents by doorstep interview and on movers by proxy interviews

Sample size reduction undermines the accuracy of the cross-sectional and longitudinal estimates.

Attrition (dropping out individuals and movers) may generate bias on estimates when those who drop out or move are systematically different from those who continue to participate. For example, if people drop out when becoming employed (change condition) because they have no more time to attend the survey, the panel produces downward biased estimates of the number of employed persons. The panel refresh, the tracking and the calibration estimator deal with the loss of sampled units.

We may reduce the negative effects of the panel attrition, by collecting some basic variables by a short form questionnaire module. The module asks for main demographic variables (age, gender, etc.), social variable (employed status, level of literacy, etc.) and, for movers, the variable on the reason of leaving the enumeration area (economic reason, personal reason, etc.). Eventually, the retrieving of phone number of the movers can favourite the phone interview for not tracked individuals.

The variable collected by the short form questionnaire module can be useful for a first step of weight adjustment for panel attrition, where the final step will be the calibration adjustment for the under-coverage. The two steps approach improves the accuracy of the estimates when the reasons of the attrition are basically different from the reasons of the sampling list under-coverage.

Doorstep interview. The doorstep interview is a short version of the main questionnaire module to obtain key information on characteristics of the non-respondents. The non-respondents are often related to target phenomena of interest. For instance, the occurrence of literacy-related non-response is obviously not random, but is thought to be largely concentrated among migrants with low literacy in the official survey languages within a country, and perhaps with low literacy in general. However, the extent of this concentration is currently unknown due to the lack of further information on these groups. Doorstep interview is a short and simple interview administered to households (or individuals) who are not willing to participate to the main survey and is aimed to retrieve this basic information for the non-respondent. As mentioned above, demographic, literacy and employment related questions are easy questions that could be administered to the non-respondent individuals. Questions 1 to 10 in figure 4.1 below are an example of this set of questions.

Proxy interview. The proxy interview can be carried out in two different contexts: one individual of the household is a mover, then another household member can answer to the short questionnaire; the whole household moves, then the neighbours (or the Community) can answer to the short questionnaire. In both cases the retrieving of the phone numbers of the movers can favourite the phone interview when tracking is not allowed by the tracking rules.

As for the doorstep interview, the task of proxy interview aims to verify the concentration of the specific characteristics in the sample of movers. A module can be administered to the proxy-respondent. Two further questions will then complement the short interviewer: that is, the telephone number of the mover and the reason for moving out of the households.

4.4. Planning the multi-mode data collection

The tracking rules are constrained by operative costs of the data collection. For such reason movers that leave the sampled area (i.e., the Enumeration Area, the Parish, District, etc.) cannot be tracked in everywhere. To retrieve the answers of movers and to reduce the panel attrition due to movers, one low-cost strategy is to perform a telephone interview. In this case, a shortened questionnaire is preferable or in more sophisticate cases, the survey plan can arrange a CATI (Computer Assisted Telephone Interview) interview and the panel survey will rely on multi-mode data collection.

Finally, the standard questionnaire has to ask the phone number to prevent a non-response of movers in the future waves.

5. EMPIRICAL EVALUATION ON THE UGANDA DATA

As a case study, we apply the estimator methods discussed above to the Uganda National Panel Survey (UNPS). The UNPS is a multi-purpose household panel survey implemented by the Uganda National Bureau of Statistics (UBOS) with the technical support of the World Bank LSMS-ISA project. Started in 2009/10 as a follow up of the Uganda National Household Survey 2005/6, its primary aim is to inform policymakers in budgetary decisions and policy interventions and for monitoring major national policies and programs such as the National Development Plan. Since its start the UNPS has been used as source of information for the compilation of the National Accounts. Implemented on an annual basis, the UNPS aims also to provide representative information on income dynamics at the household level and provide information on consumption expenditure estimates to monitor poverty in interval years of other national survey efforts. Moreover, the UNPS collect high-quality agricultural data integrated within a multi-topic framework and, thus, allows for understanding linkages between agriculture and welfare, and other socio-demographic characteristics.

Since its launch in 2009, the UNPS conducted 8 waves of data collection on a sample of about 3,000 households. The UNPS 2009/10 was, indeed, followed by additional rounds in 2010/11, 2011/12, 2013/14, 2015/16, 2017/18, 2018/19 and 2019/20. The samples are representative at the national, rural-urban and regional level. This study uses data from the UNPS 2009/10 (wave 1), the UNPS 2013/14 (wave 4), and the UNPS 2015/16 (wave 5). The UNPS 2009/10 comprises a sample of 2,975 households; the UNPS 2013/14 counts 3118 households; in the UNPS 2015/16, 3,304 households had a complete interview. After some data cleaning and data preparation for the analysis, the final dataset comprises 18,313 individuals from wave 1; 17,377 individuals from wave 4 and 15,905 individuals from wave 5.

To illustrate the dynamics in the three Survey waves analysed, table 6.1 shows the attrition computed as the share of lost observation in the 3 waves of data collection used in this work. Moreover, it includes, the number of individuals entering each year with respect to the baseline year of the survey. Finally, it presents the dynamics separately for the original sample portion and for the part of the sample rotated in starting in 2013/14. The 60 per cent of individuals sampled in 2009 are no longer part of the sample in 2015/16. Most individuals were dropped out from the sample in 2013/14 also as consequence of the refresh occurred in that wave. In 2013/14, 10,102 individuals are left out of the sample, while 9,166 entered for the first time. 5,387 individuals enter through the sample refresh. Indeed, as show in column four in the table, 3,779 individuals are new individuals in the original sample of households. The analysis in this work, will focus on the 2009/10-2015/16 and 2013/14-2015/16 panels. In each of the panels, only the individuals present in both years are considered in the analysis. Since there are not

individuals entering again the sample in wave 5 after leaving in wave 3, the final number of observations for the first of the two panels is 7,215; the second counts 4,688 total units.

Table 6.1. Attrition and new entries in the UNPS 2009/10 - 2013/14 - 2015/16 panel; in the UNPS 2009/10 - 2013/14 original sample; and in the UNPS 2013/14 - 2015/16 rotated sample

.	UNPS 2009/10 ; 2013/14 ; 2015/16		UNPS 2009/10; 2013/14 (original sample only)		UNPS 2013/14 ; 2015/16 (rotated sample only)	
	Individuals as share of the initial sample (%)	Individuals (N)	Individuals as share of the initial sample (%)	Individuals (N)	Individuals as share of the initial sample (%)	Individuals (N)
Initial Sample	**100**	**18,313**	**100**	**18,313**	**100**	**5,387**
Individuals in all years (balanced panel)	**39.4**	**7,215**	**44.8**	**8,211**	**87.0**	**4,688**
Total individuals who left	**60.6**	**11,098**	**55.2**	**10,102**	**13.0**	**699**
Individuals who left before 2013/14	55.2	10,102				
Individuals who left before 2015/16	5.4	996				
Total new entries	_	**10,325**	_		_	
Individuals who entered in 2013/14	_	9,166	_	3,779	_	
Individuals who entered in 2015	_	1,159	_		_	139

Table 6.1 is the results of complex protocol defining the tracking rules. The UNPS tracking scheme considers the mobility of the target population by following i) those households moving away from their initial location as a whole; ii) those households who shifted with some members to another location while the other members went elsewhere and became split-offs, and iii) individuals moving out their original household to form new split-off households. Tracking rules have changed over the course of the Panel rounds for all three tracking targets mentioned above.

In the first three waves, the UNPS tracked all the original households that shifted away from their 2005/06 original location to any other location within or outside the same EA, although not all households of the initial sample were targeted for individual tracking, in case any of their members moved out of the household. Indeed, the tracking of the individuals and therefore of the split-offs they form, was covering only a subsample of the original households, comprising 2 households randomly selected per EA, that is the 20 per cent of the sample of the original households. This individual tracking was meant to compensate the losses to the sample due to the attrition. These 20 per cent of households from which the individuals were tracked, were referred to as split-offs tracking targets. Moreover, not all the individuals of the split-offs tracking targets were eligible for tracking. Indeed, only members of 15 years and above and biologically related to the household head were tracked. Individuals and split-off households were found and interview even if they moved beyond their original EA/parish.

Once interviewed, the split-off individuals and all the members of the new household they formed or joined became part of the UNPS sample and were interviewed and eventually tracked in all subsequent UNPS waves.

Starting in UNPS 2013/14 the scope of the split-off target tracking was expanded to include all households part of the sample, regardless of the fact they were original or split-offs, and of location or distance from original household location provided they were still residing in

Uganda. In those households, only members older than 15 years and identified in the previous wave as the head, the spouse or the child of the household head were eligible for tracking. Other members were interviewed only if leaving with one of these members. This means that if no core member is found in the last known location, the household was not interviewed even if other previous household members still lived there.

In wave 4 the sample of the UNPS was refreshed and one third of it was rotated out. This one-third of the original sampled household rotated out as part of the panel refresh was no longer tracked or interviewed at all.

In the current UNPS setting, the tracking of individuals entails the completion of an individual tracking form comprising all the contact information of the split-offs or the individual movers. The information on their new location needed for the full tracking is generally gathered from their previous household members of any other knowledgeable person. For each core member that had moved away, a tracking form was completed. Based on the information filled in this form, the mover individuals are contacted, and then interviewed.

Although the tracking target sample comprises only the core members of each household, when they move, all persons living with them are interviewed and become part of the UNPS sample. Finally, if these persons are core members of the new split-off household, they are interviewed in the subsequent waves of the UNPS, even if they shift to different locations.

5.1. Empirical evaluation of the cross-sectional estimates

We computed the UNPS 2015/16 estimates. Data from the previous waves are included in the analysis to identify the household dynamics (e.g. movers, immigrants and new-borns).

We calibrated the base weights to the known sex by age class population totals using UBOS official projections for 2015 and based on the Population Census of 2014. The projections used to calibrate the weights are presented in the table 5.2 below.

Hereinafter, we denote these weights as calibrated GWSM base weights.

As far the practical implementation of this experiment, we can give the following indications:

- We designed the new estimation process to change as little as possible the current process.
- Uganda panel data stores the variables, the *personal identification code*, the *household identification code in the original sample* and the *household identification code in the current wave*. The variable that identifies the *original selection sample is not explicitly stored since it does not clearly distinguish* the non-respondent in the original sample (first wave) with a new individual that is not included in the original sample.
- We artificially created the population membership variable, $F_{k,t\leftarrow t^*}$. Using other stored variables, making some accurate assumptions. The variable should be directly asked in the new version of the questionnaire.
- We artificially created the *individual multiplicity factor* using other stored variables, making accurate assumptions. The variable should be directly asked in the new version of the questionnaire (Does the individual of the original sample change Parish from 2009 to 2013 (years of the actual samples).
- Based on the values of the *individual multiplicity*, we computed the household multiplicity factor.

- We artificially rebuilt the response indicator variables in each wave (= 1 for respondents, and = 0 non-respondents or not in the original sample) using other stored variables and making accurate assumptions.
- We stored the inclusion probability of each individual selected in and original sample (not for the incorporated individuals) even though the individual does not respond.
- We computed the GWSM weight that is not relied on the concept of the parent household which can be a complex operation to be done in practice;
- We computed the final weights using the standard calibration estimator. The final weights are applied for the cross-sectional and longitudinal estimates. That means we have a unique vector of sampling weights simplifying the coherence of the individual and household estimates.

Table 6.2. Population projection by age and gender (2015)

Age group	Male	Female
0-4	3,220,300	3,028,800
5-9	2,870,800	2,730,500
10-14	2,528,000	2,500,100
15-19	2,008,700	2,097,500
20-24	1,509,200	1,790,100
25-29	1,181,200	1,403,800
30-34	938,700	1,082,800
35-39	743,900	836,500
40-44	630,600	677,500
45-49	470,800	486,800
50-54	382,300	443,200
55-59	241,700	277,900
60-64	194,600	241,600
65-69	140,400	171,100
70-74	114,900	158,600
75-79	71,500	90,700
80+	96,400	140,600
Total	17,344,000	18,158,100

We have applied the calibrated GWSM base weights to a set of variables from UNPS 2015/16 data and compared them to Uganda official statistics to assess the functioning of the weights vis-à-vis the UNPS 2015/16 sampling weights. We use published data and reports from the National Population and Housing Census (NPHC) 2014 (UBOS 2016, UBOS 2017) and the Uganda National Household Survey (UNHS) 2016/17 as main sources of official statistics.

While the former is conducted by UBOS about every 10 years with the aim of collecting bench-mark demographic and socio-economic data of the Uganda population. The latter is the sixth follow-up survey of the UNHS, a cross-sectional survey implemented by UBOS starting in 1999/20. It aims to collect socio-economic data on demographic and socioeconomic characteristics. Its sample counts 15,636 households. The UNPS is a follow-up to its 2005/06 survey and implements in large extent the same methodology.

The NPHC 2014 and UNHS 2016/17 are the official source of data with the closest collection period to the UNPS 2015/16 we found: while the reference period for the UNPS 2015/16 is March 2015 to March 2016; data collection for NPHC 2014 was pursued in August/September, whereas the UNHS 2016/17 data collection was implemented between June 2016 and June 2017.

Table 6.3 presents the comparison estimate indicators on individual and household characteristics at the national level. The first three columns show the indicators from the UNPS 2015/16 without using sampling weights (unweighted estimator), using the original UNPS weights (current UNPS estimator) and using the calibrated GWSM base weights (calibrated GWSM base estimator). In the last three columns, the table presents the official statistics and their source. In general terms, estimates are more accurate using the calibrated GWSM base estimator rather than the current UNPS estimator for indicators at the individual level. The share of female population and the share of children below 18-years-old weighted using the calibrated base GWSM weights approximate well the official statistics. This result is attended since the calibrated weights reported to the projection of the population totals for sex and age classes for 2015. Projections that are based on the 2014 Census that we are using as benchmark.

Regardless the weight we apply, the UNPS 2015/16 overestimates the official "true" value of the literate population above 10 years. However, the three estimates are consistent within each other. The unweighted statistic with much higher value (80.6%) than the official statistics (72.2%) suggests that the nonresponse affects the not literate population and for such reason this sub-population is under-represented in the panel. The calibration of the proposed estimator is not able to deal with this problem. Finally, the calibrated GWSM base weight estimator appears to be more accurate for two key indicators of the survey, that is the Working population and the poverty head count at the national level.

At the household level, the current UNPS and the calibrated GWSM base estimators seem more comparable, even though the former produces more accurate statistics especially for the share of urban households and share of agricultural households variables. In few cases – namely, the access to electricity and the ownership of the dwelling – the UNPS 2019/20 produces inaccurate estimates of the official statistics regardless of the weight used. The value of the unweighted estimators for these two indicators suggests for further investigation on discrepancies on the methods of data collection.

The findings on the households advise us to implement a calibration involving household known totals.

Table 6.3. Individual and household estimates by the unweighted, current UNPS and calibrated GWSM base estimators

	Unweighted	UNPS	Calibrated GWSM base	Official Statistics	
	Mean	Mean	Mean	Mean	Source
Individuals					
Female	51.3	50.8	51.0	51.0	Census (2014)
% children below 18 years old	48.9	51.4	55.1	55.0	Census (2014)
Literate population (+10 yrs)	80.6	80.8	81.1	72.2	Census (2014)
Working population	76.7	76.1	80.0	78.8	UNHS (2016)
Poverty headcount	19.7	19.0	20.0	21.4	UNHS (2016)
Households					
Share of urban households	25.0	25.1	21.4	25.0	Census (2014)
Size of the Household	4.8	5.0	5.0	4.7	Census (2014)
Share of Agricultural Household	79.0	80.2	81.5	80.0	Census (2014)
Number of rooms in the Household	2.2	2.3	2.1	2.4	UNHS (2016)
Household owns a Cellphone	74.1	72.9	74.9	74.3	UNHS (2016)
Household has access to electricity	15.4	15.1	14.4	21.0	Census (2014)
Household has access to safewater	75.2	72.2	71.9	72.0	Census (2014)
Household owns the dwelling	82.1	83.0	82.9	71.8	UNHS (2016)

House has brickwalls	67.3	63.8	63.9	66.6	UNHS (2016)

Table 6.4 breaks down the share of working population by area of residence. The calibrated GWSM base estimator (82.0%) seems more accurate than the current UNPS estimates (78.8%) in representing the official statistics in rural area (82.6%). More controversial are the estimates in the urban areas where we do not see a best estimator.

Table 6.4 . Employed population by area of residence

	Unweighted	UNPS	Calibrated GWSM base	Official Statistics		
	Mean	Mean	Mean	Mean	Year	Source
Rural	79.6	78.8	82.0	82.6	2016	UNHS
Urban	67.5	67.3	71.8	69.0	2016	UNHS

The focus on poverty by area of residence (table 6.5) does not show concrete differences among the estimators. The incidence of poverty is much higher in the rural areas than urban areas.

Table 6.5. Poverty head-count ratio by area of residence

	Unweighted	UNPS	Calibrated GWSM base	Official Statistics		
	Mean	Mean	Mean	Mean	Year	Source
Rural	23.6	21.8	22.4	25.0	2016	UNHS
Urban	9.0	9.7	10.0	9.6	2016	UNHS

5.2. Empirical evaluation of the longitudinal estimates

The panel survey produces net-change and gross-change estimates (Section 3.2.2). Gross change gives some valuable insights into the net change so it is important that the cross-sectional estimates achieved by transition matrices are consistent with the estimates obtained with the overall UNPS 2015/16 sample.

The table 6.6 shows the dynamics of employment across the 2009 and 2015, and 2013 and 2015 panel. In particular, the first three columns reports the employment status in 2015 respect to the status in 2009. Only individuals present in both 2009 and 2015 are represented in the table. The last three column show the same statistics for those individuals in 2015 who responded to the survey also in 2013.

Let us focus on the 2009/2015 estimates. We note all the estimators produce upward estimates with respect to the estimates of table 5.3. This result suggests the individuals who left across the years – and characterizing most part of the attrition – were mostly unemployed. The three estimators gave respectively the 78.2%, 77.5% and 80.7% of employed persons. The refresh operation rebalanced the panel. Focusing on the cross-sectional estimates given by the 2013/2015, the employed persons by the unweighted estimator are 77.4%, for the current UNPS estimator are 76.6% while for the calibrated GWSM base weight estimator are 80.7%.

We can state that the last estimator produce more stable and consistent estimates when considering different sub-samples, i.e. the 2009/2015 panel (80.7%), the 2013/2015 panel (80.7%) and the entire sample observed in the 2015 (80.0%).

Table 6.6. Employment transition matrix estimates by the unweighted, current UNPS and calibrated GWSM base estimators

	2009/15			2013/15		
	Unweighted	UNPS	Calibrated GWSM base	Unweighted	UNPS	Calibrated GWSM base
Unemployed in 2009 and 2015	5.2	5.4	4.0	11.6	12.1	9.5
Unemployed in 2009 and employed in 2015	5.9	5.6	5.8	8.5	8.1	7.4
Employed in 2009 and 2015	72.2	71.9	75.0	69.0	68.5	73.2
Employed in 2009 and unemployed in 2015	16.6	17.1	15.2	11.0	11.2	9.8

The breakdown of employment dynamics in the 2009 and 2015, and 2013 and 2015 panels by area of residence shown in table 6.7 helps to understand that the panel attrition is concentrated on the unemployed in the urban areas. Comparing the values of table 6.4 in the rural areas the estimates are consistent with the ones computed in the table 6.7. The refresh (2013/2015) improves the accuracy. Instead, when considering the urban areas all the methods gave upward estimates especially with the sample 2009/2015. The findings indicate that individuals leaving the UNPS are mostly unemployed in urban areas and suggest improving the tracking rules to reduce the potential bias of the estimates.

Table 6.7. Employment by area of residence transition matrix estimates by the unweighted, current UNPS and calibrated GWSM base estimators

		2009/15			2013/15		
		Unweighted	UNPS	Calibrated GWSM base	Unweighted	UNPS	Calibrated GWSM base
Unemployed in 2009 (or 2013) and 2015	Rural	3.9	4.2	3.3	8.9	9.2	7.5
	Urban	9.8	9.4	7.6	20.4	21.3	17.3
Unemployed in 2009 (or 2013) and employed in 2015	Rural	4.3	3.9	4.4	8.0	7.7	6.9
	Urban	11.7	11.4	12.9	10.0	9.4	9.4
Employed in 2009 (or 2013) and 2015	Rural	75.0	74.4	76.7	71.7	71.2	75.3
	Urban	62.2	63.3	66.1	59.5	59.9	65.0
Employed in 2009 (or 2013) and unemployed in 2015	Rural	16.7	17.4	15.6	11.3	11.8	10.2
	Urban	16.3	15.9	13.4	10.1	9.3	8.2

We carried out similar analysis for the poverty head-count ratio (table 6.8). The 2009/2015 over-counted the poverty. For the three estimators we have respectively 21.8%, 21.0% and for the calibrated GWSM base estimator 22.4%. The refresh operation (2013/2015 sample) reduce the frequencies at 19.8%, 19.1% and 20.1%. When we compare these estimates with the official statistic (21.4%) it does not appear that one estimator overcomes the other.

Table 6.8. Poverty rate transition matrix estimates by the unweighted, current UNPS and calibrated GWSM base estimators

	2009/15			2013/15		
	Unweighted	UNPS	Calibrated GWSM base	Unweighted	UNPS	Calibrated GWSM base
In non-poor HH in 2009 (or 2013) and 2015	63.7	65.2	63.3	65.3	66.4	64.9
In non-poor HH in 2009 (or 2013) and poor HH 2015	12.3	12.0	13.0	8.0	8.0	8.4
In poor HH in 2009 (or 2013) and poor HH 2015	9.5	8.9	9.3	11.8	11.1	11.8
In poor HH in 2009 (or 2013) and non-poor HH 2015	14.5	13.8	14.4	14.8	14.3	14.9

Table 6.9 shows the transition matrix considering the area of residence. For both the area of residence the 2009/2015 panel the three estimates are respectively greater than 2013/2015 estimates. Considering the estimated using the complete sample (table 6.4) the calibrated estimates (2009/2015 and 2013/2015) appear more stable. The overestimations of poor people suggest that the non-poor individuals, especially in rural areas, are more dynamic and tends to move out of the UNPS sample more than poor individuals. In the 20013/2015 panels figures stabilize. All statistics are consistent with the National level estimates from the UNPS 2015 cross-sectional sample. The refresh of the sample in 2013/14 seems to have mitigate the attrition of non-poor individuals. Rural and urban area of residence shown the same trends.

Table 6.9. Poverty rate by area of residence transition matrix estimates by the unweighted, current UNPS and calibrated GWSM base estimators

		2009/15			2013/15		
		Unweighted	UNPS	Calibrated GWSM base	Unweighted	UNPS	Calibrated GWSM base
In non-poor HH in 2009 (or 2013) and 2015	Rural	58.3	60.1	59.8	59.4	61.5	60.7
	Urban	81.4	83.4	81.4	84.2	83.6	83.1
In non-poor HH in 2009 (or 2013) and in poor HH in 2015	Rural	14.1	13.6	14.1	9.1	8.7	9.3
	Urban	7.0	6.5	7.4	5.4	5.5	4.5
In poor HH in 2009 (or 2013) and 2015	Rural	11.5	10.5	10. 7	14.6	13.2	13.2
	Urban	3.2	3.5	2.4	3.9	4.2	5.7
In poor HH in 2009 (or 2013) and non-poor HH in 2015	Rural	16.1	15.8	15.4	16.9	16.6	16.9
	Urban	8.3	6.7	8.8	6.5	6.6	6.7

6. CONCLUSIONS

This paper addressed the quality of panel surveys considering issues related to the sampling design, estimation and data collection.

We have seen that an improvement in the quality of investigations passes precisely through a strategy that takes the three aspects mentioned above together.

The strategy proposed is feasible, has a limited impact on the current survey practices and enhances the accuracy of the final estimates.

Moreover, we carried out an empirical experiment on the Uganda National Panel Survey comparing the sampling estimates of the new methodology with those given by the current one. We can summarize the results as follows:

- the calibrated GWSM base estimator seems to produce more accurate individual statistics than the current UNPS estimator;
- the two estimators produce equally accurate statistics at household level. That suggests to improve the calibrate estimator accounting available household known totals;
- the cross-sectional estimates based on the transition matrix generally do not reproduce exactly the cross-sectional estimates based on the entire UNPS sample. The refresh operation mitigates the problem. This suggests: i) to implement coherently a periodic sample refresh (rotate panel); ii) to improve the tracking rules since the panel attrition affects specific sub-populations of interest;
- the calibrated GWSM base cross-sectional estimates on the transition matrix appears generally more stable when changing the sample (i.e 2009/2015 or 2013/2015) than the current UPNS estimates.

REFERENCES

Deville, J. C. and Särndal C.E. (1992). Calibration Estimator in Survey Sampling. *Journal of American Statistical Association*. Vol.87. N° 418.

FAO (2015). *Integrated Survey Framework*, Guidelines, Rome. http://www.gsars.org/wp-content/uploads/2015/05/ISF-Guidelines_12_05_2015-WEB.pdf. Accessed on 16 December 2015.

FAO (2014). *The Global Strategy to Improve Agricultural and Rural Statistics. Technical Report on the Integrated Survey Framework*, Technical Report Series GO-02-2014, http://gsars.org/wp-content/uploads/2014/07/Technical_report_on-ISF-Final.pdf. Accessed on 1 December 2014.

Groves, R.M., Don A. Dillman, J. L. Eltinge, R- J. A. Little. (2001). *Survey No-Response*, Wiley. New York.

Falorsi, F,, Russo A. (2009). Popolazioni e parametri di interesse nelle indagini ripetute nel tempo. In Alleva G., Falorsi P. *Indicatori e modelli statistici per la valutazione degli squilibri territoriali*. Franco Angeli, Milano. ISBN: 9788856812077.

Horvitz, D.G. and D.L. Thompson. (1952). A generalisation of sampling without replacement from finite-universe. *J. Amer. Statist. Assoc.*, 47,663-685.

Lavallée, P. (2007). *Indirect Sampling*. Springer: Ottawa.

Lynn, P., (2009), *Methodology of Panel Surveys,* Wyley. Wiley and Sons, Ltd, Publication.

Mecatti, F. (2007). A single frame multiplicity estimator for multiple frame surveys. *Survey Methodology*, 33(2): 151-157.

Narain, R.D. 1951. On sampling without replacement with varying probabilities. *J. Ind. Soc. Agril. Statist., 3,169-174.*

Singh, A.C. & 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.

Uganda Bureau of Statistics (2017), The National Population and Housing Census 2014 – Education in the Thematic Report Series, Kampala, Uganda.

Uganda Bureau of Statistics (2016), The National Population and Housing Census 2014 – Main Report, Kampala, Uganda.

University of Essex. (2016), *British Household Panel Survey*, https://www.iser.essex.ac.uk/bhps

World Bank. (2020), *Basic Information Document - The Uganda National Panel Survey (UNPS) 2015/2016*. http://surveys.worldbank.org/lsms/programs/integrated-surveys-agriculture-ISA/uganda#bootstrap-panel.