# LONGITUDINAL DATA FOR THE ANALYSIS OF ITALIAN LABOUR MARKET FLOWS

**Barbara Boschetto, Antonio R. Discenza, Carlo Lucarelli, Simona Rosati** [1]

*Italian National Institute of Statistics, Rome, Italy*

**Francesca Fiori**[2]

*Centre for Population Change, University of St. Andrews, St. Andrews, United Kingdom*

***Abstract.*** *The Italian Labour Force Survey belongs to the framework of the European Union Quarterly Labour Force Survey and satisfies completely all the EU Regulations. It has a 2-2-2 rotating sample, thus, given a quarterly sample, 50% of the households are interviewed again after 3 and 12 months. By matching the records associated to the same individual for the different quarters, it is possible to build longitudinal datasets and transition matrices. Longitudinal data allow an accurate evaluation of the level of labour market mobility, in particular at a time of economic recession when dynamics result markedly accentuated. This paper, thus, presents the complete framework to produce gross flows estimates consistent with the quarterly estimates already disseminated.*

***Keywords:*** *Labour Force Survey, Longitudinal data, Weighting, Calibration.*

## 1. INTRODUCTION

Since 1959, in Italy, the Labour Force Survey (henceforth LFS) represents the main source of information on the labour market (Istat, 2006); quarterly, it provides estimates of the absolute values of employment, unemployment and inactivity, and of both their trends and conjunctural variations. Such estimates refer to all private households residents. Residents of communal establishments (hospices, orphanages, religious institutions, etc…) are not surveyed.

The survey includes a longitudinal component which derives from the rotational scheme of the sampled households. Archives of longitudinal micro-data may be obtained by linking the information collected on both the same individual and household at different quarters. The transition matrices obtained from the

[1] Barbara Boschetto: boschett@istat.it, Antonio R. Discenza: discenza@istat.it, Carlo Lucarelli: calucare@istat.it, Simona Rosati: sirosati@istat.it

[2] Francesca Fiori: ff20@st-andrews.ac.uk.

longitudinal dataset provide an estimate of the number of permanencies within and transitions between the various occupational conditions, allowing a deeper understanding of labour market dynamics and of the involved individuals.

The longitudinal micro-data files and the transition matrices constitute a "by-product of the survey, which is not a proper panel referred to the whole population (Tate and Clarke, 1999). As a matter of fact, if an individual - originally interviewed in one of the sampled municipality - changes his residence between the first and the subsequent interview, he will not be interviewed again.

This paper presents the most relevant aspects concerned with the construction of 12 months longitudinal datasets and of associated transition matrices of gross flows. Although the Labour Force Survey is conducted in all EU and Candidate countries, and its implementation is strictly disciplined by Council Regulations (European Commission, 1998), at the current time not all the states are involved in the production of longitudinal datasets. Furthermore, the sample design and rotation patterns are not fully harmonised. As a consequence, the national statistical institutes do not follow a standard procedure in the production of longitudinal datasets. This paper describes the original approach developed by the Italian National Institute of Statistics.

The relevance of the paper is twofold. First, the methodology described may represent a valid and reliable source of reference for the production of longitudinal micro-data from the cross-sectional LFS within the wider EU context. More generally it may contribute to the theoretical and methodological debate around the importance of exploiting all the available information by linking together information from existing sources. Secondly, given the relevance of flow estimates to assess labour market dynamics – particularly in this period characterised by a strong downturn – this paper may constitute an instrument to academics and researchers interested in the field. The paper, in fact, may be seen as companion guide to the longitudinal datasets, which illustrates the structure of the included information, the methodology applied to obtain the final datasets, and the meaning of the figures that can be derived, thus allowing a correct and critical exploitation of the data-source**.**

The adoption of a longitudinal perspective allows assessing transition and persistence probabilities for "employment" and "unemployment".

Transition matrices containing estimates of gross flows can be produced provided that the following aspects are taken into account:

a) LFS is not a full panel survey; thus, individuals who move out of the selected households, or households which move out of the municipality, are not re-interviewed;

b) household non-response may occur at subsequent waves due to non-contact, refusal, etc.;
c) longitudinal estimates can refer only to a specific longitudinal reference population. This is defined in Italy as the population which reside in the same municipality for the entire period under consideration (12 months) thus net of deaths and internal or international migration;
d) weights should reflect the longitudinal population, accounting for the total non-response (usually not at random).

In the next sections, the paper will address the most salient issues involved in the production of the longitudinal datasets, and of the corresponding transition matrices, following the structure below:

- record linkage of individuals (Section 2);
- longitudinal data editing and imputation (Section 3);
- estimation of the reference longitudinal population (Section 4);
- weighting procedure and treatment of non-response (Section 5) and
- concluding remarks (Section 6).

## 2. REKORD LINKAGE OF INDIVIDUALS

### 2.1 SAMPLE DESIGN AND THE LONGITUDINAL COMPONENT

Since the first quarter 2004, the LFS underwent a profound restructuring, but the sample design remained almost unvaried. The sample is based on a two-stage design, with stratification of the first-stage units (municipalities) and rotation of second-stage units (households) with a 2-2-2 scheme.

Samples referring to distinct quarters, thus, partially overlap. Each household is included in the sample for two consecutive occasions and, after a pause of two quarters, is then included again in the sample for two other occasions. Fig. 1 reports the distribution of rotation groups with regards to different waves of the survey. The sample of households interviewed in the first quarter 2008, for instance, consists of a group of households entering the sample for the first time (F1), a group of households at their second interview (E2), another group at their third interview (B3) and, lastly, a group of households which are at their forth and last interview (A4).

Specifically, every quarter, the theoretical sample is composed of around 76 thousand households. These are equally divided among the four rotation groups, so that each group consists of nearly 19 thousands households.

| REFERENCE PERIOD | | ROTATION GROUP | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Quarter 4 | 2006 | A1 | | | | | | | | |
| Quarter 1 | 2007 | A2 | B1 | | | | | | | |
| Quarter 2 | 2007 | | B2 | C1 | | | | | | |
| Quarter 3 | 2007 | | | C2 | D1 | | | | | |
| Quarter 4 | 2007 | A3 | | | D2 | E1 | | | | |
| Quarter 1 | 2008 | A3 | B3 | | | E2 | F1 | | | |
| Quarter 2 | 2008 | | B4 | C3 | | | F2 | G1 | | |
| Quarter 3 | 2008 | | | C4 | D3 | | | G2 | H1 | |
| Quarter 4 | 2008 | | | | D4 | E3 | | | H2 | I1 |

**Figure 1: Rotation scheme of the sampled households**

A similar longitudinal framework implies a 50% overlap of the theoretical sample over 3 and 12 months, and a 25% overlap over 9 and 15 months. This partial overlapping allows the construction of longitudinal micro-data files over 3, 12 and 15 months. For instance, the 12 months longitudinal dataset, referred to the first quarter 2007 and the first quarter 2008, contains information on households included simultaneously in the groups B1 – B3 and A2 – A4 (Fig. 1). The level of estimates precision is inevitably reduced compared to the quarterly cross-sectional survey. Therefore, figures referred to small areas or population sub-groups may be affected by higher levels of uncertainty.

### 2.2 RECORD LINKAGE

The implementation of the record linkage procedure finds inspiration from the work of Fellegi and Sunter (1969) and other works that have developed more recently this aspect as Winkler (1995), Torelli (1998) and Paggiaro and Torelli (1999). 50% of the sample, consisting of two different groups of rotation (i.e. groups A and B), is interviewed again after twelve months (Fig. 1). Although the purpose is to obtain labour market flows referred to twelve months intervals, we chose to link information from all four waves of survey for each rotation group of the quarterly sample. In doing so, the entire available information is used during the process of data editing, ensuring maximum consistency. More precisely, we first

identified the reference population, by linking individuals who were interviewed in two waves at twelve months of distance. In a second stage, information on the two other waves (at three and fifteen months) was also attached.

The presence of a unique individual identification code within the family (individual-key, composed of six digits), made the record linkage quite straightforward, although it may sometimes be affected by errors. Therefore we also relied on additional information derived from the following personal identifiers: date of birth, sex and name.

A file of *"matchable"* individuals (individuals who could potentially be linked) was first obtained, by selecting all the households that were interviewed in both matched quarters. In this file couples of individuals linked through individual-key were distinguished from non-linked individuals.

A deterministic procedure was applied to verify the coincidence of the individual keys and personal identifiers (name, sex and date of birth). Then, a specific code was attributed to the individuals according to the extent of matched information.

The application of this deterministic procedure allowed identifying first of all the individuals who were correctly matched, i.e. who showed perfect coincidence of individual key and personal identifiers. Furthermore, the procedure singled out the "false negatives" (when the same person has two different individual-keys), and the "false positives" (when a single individual-key is associated with two different people).

In this way individuals who were matched correctly were immediately identified, because after a year they had the same individual-key and the same personal identifiers.

Individual-key affected by error is a possible event although very rare, and it is often related to the swap of individual key within the same household. Thus the individuals who were not matched by individual-key or who had discordant personal data in two quarters, were compared between waves to other individuals within the same household, through the only personal identifiers (date of birth, sex and name). Individuals who showed a perfect coincidence of all variables (except individual-key) were then recovered.

The initial dataset of *"matchable"* individuals (composed of 79,151 individuals in households interviewed in the first quarter of 2007 and 2008 could then be subdivided as shown in Table 1.

Although this procedure is not automatic, it has the advantage of being divided into distinct and replicable phases, so that it may be applied for record linkage of other datasets.

**Table 1: Record linkage rates**

| **Individuals in households interviewed in the first quarter of 2007 and 2008** | **n.** | **%** |
|---|---|---|
| **Matchable records** | **79,151** | **100.0** |
| **- Linked records** | **76,985** | **97.3** |
| -- Records correctly matched | 75,854 | 95.8 |
| -- Records with wrong individual-key | 1,131 | 1.4 |
| ---- Linked records recovered from false positives with exchange of key | 740 | 0.9 |
| ---- Unlinked records recovered from false negatives with exchange of key | 391 | 0.5 |
| **- Total non-matched records** | **2,166** | **2.7** |
| -- False positives | 567 | 0.7 |
| -- Individual records that are present in a single wave of the period for analysis | 1,599 | 2.0 |

To complete the reconstruction of the longitudinal file, the same record linkage procedure was used to attach the information from the other waves. By doing so, information for each rotation group was obtained and then subsequently used in the editing and imputation phase. In the example provided in this paper, the sample of matched individuals after twelve months is composed of two rotation groups: A and B.

Thus, matched data for the period 1st quarter 2007-1st quarter 2008 come from the 2nd and 4th wave of interview for the rotation group A, and from the 1st and 3rd wave of interview for the rotation group B.

Then, the other two waves of interview attached to each rotation group come from distinct quarters. Specifically, to the rotation group A we attached data from the 1st interview, which took place in the 4th quarter of 2006 and from the 3rd interview, which took place in the 4th quarter of 2007. To the group B we attached data from the 2nd interview, which took place in the 2nd quarter of 2007, and from 4th interview, which took place in the 2nd quarter of 2008 (Rosati and Boschetto, 2013).

After the application of the record linkage procedure, two separate longitudinal data files with the information from each time of the survey were obtained, one for each rotation group. These files were then used separately in the editing process, and finally joined again for analysis of all individuals interviewed in the first quarter of 2007 and in the first quarter of 2008.

## 3. EDITING AND IMPUTATION

This paragraph is concerned with editing and imputation of wave-to-wave inconsistencies in the LFS longitudinal data. Such errors, which refer to data changing between two different survey waves, may be related to various sources of non-sampling errors such as panel conditioning, telescoping effect or recall errors (an overview of different types of error in longitudinal survey is provided by Lynn *et al*., 2005).

Handling wave non-response, which occurs when responses are obtained for some but not all waves of the survey, is out of scope of our study. See Kalton (1986) and Lepkowski (1989) for a review of the issues involved in compensating for wave non-response in panel surveys.

As a result of the linkage process, panel members can be divided into three groups of permissible patterns: complete respondents, one-wave non-respondents, two-wave non respondents. As showed in the Table 2, a substantial number of panel members responded in all the four waves of survey (the proportions for the groups A and B were found to be 81.3 and 85.9, respectively).

**Table 2: Response patterns in four waves for the rotation groups A and B**

| Response status | wave 1 | wave 2 | wave 3 | wave 4 | n. | % |
|---|---|---|---|---|---|---|
| Complete | A1 | A2 | A3 | A4 | 31,936 | 81.3 |
| One wave non-respondents | A1 | A2 | —- | A4 | 2,589 | 6.6 |
| | —- | A2 | A3 | A4 | 4,058 | 10.3 |
| Two wave non-respondents | —- | A2 | —- | A4 | 704 | 1.8 |
| Complete | B1 | B2 | B3 | B4 | 32,413 | 85.9 |
| One wave non-respondents | B1 | B2 | B3 | —- | 3,116 | 8.3 |
| | B1 | —- | B3 | B4 | 1,050 | 2.8 |
| Two wave non-respondents | B1 | —- | B3 | —- | 1,119 | 3.0 |

— = non-respondent

Units that failed to provide data for either the 1st quarter of 2007 or the 1st quarter of 2008, i.e. units for which longitudinal information is missing, are not included in the process.

The choice of an appropriate method for longitudinal imputation is not straightforward. Relying on previous experiences, we know that several considerations must be made before selecting an imputation method (Rosati, 2004). For our purpose, we decided to develop a longitudinal deterministic method for

several reasons. First, the type of respondent, self or proxy, can be used as auxiliary variable for assessing reliability of inconsistent responses across the waves. Secondly, given that CATI is a dependent interviewing method, the variable related to previous information was also incorporated as auxiliary variable. Furthermore, the entire work history for each individual, from the first wave to the fourth, was investigated in order to derive general deterministic rules of longitudinal imputation. These rules allowed deducing the imputed value from data available in either current or previous wave on the basis of the relation between the auxiliary variable and the corresponding item non-response.

More precisely, the imputation strategy provides that the errors related to the first two waves can be corrected by applying an algorithm which imputes the incorrect variable changing the value either in the first wave or in the second wave according to hierarchical rules (Rosati and Boschetto, 2013).

In the subsequent waves, longitudinal imputation consists of using data from the previous wave to impute inconsistent data in the current wave (*conditional imputation*). Although this rule adds further constraints to the entire process, it ensures a basic requirement for the LFS that is to obtain final estimates that do not need to be revised when a new wave is available.

## 4. ESTIMATION OF THE LONGITUDINAL POPULATION

Labour market flow figures could in principle provide estimates of all flows regarding the population living in a country at the beginning of the period of observation. Or they could also provide estimates of all the gross flows which have determined net changes between two cross-sectional estimates.

In the first case, longitudinal data from LFS should ideally represent the whole initial population. However, the initial population modifies during the period of observation due to demographic events such as deaths and internal/international migration. Thus, longitudinal data could represent the whole initial population only if the LFS was designed as a "proper" panel, in which all the individuals in the initial sample were "followed" over time and re-interviewed. If this was the case, information would be collected also on individuals moving to another municipality or to another country. And we would be able to identify deaths of sample individuals. However, since the LFS is not designed as a panel, we cannot gather all the information that would be ideally desirable. In the second case, we should bear in mind that annual or quarterly net changes are the final result of a high number of gross flows (Fig. 2) of different nature and different magnitude (death, migration, labour status transitions).

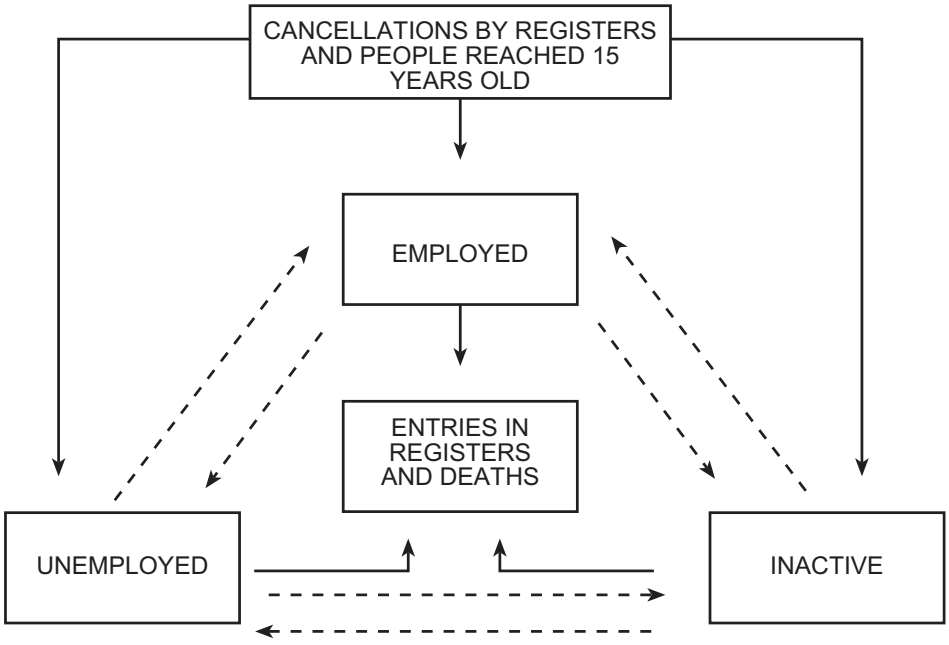


**Figure 2: Gross flows of population on labour market**

Given the current design of the Italian LFS (and of the EU-LFS in general), therefore, we could only expect to derive transition matrices like that displayed in Fig. 3 where:

- T is the transition matrix containing stock and flow estimates for the total longitudinal population (i.e. still resident in the country at the end of the period);
- S1 is the vector containing official cross-sectional estimates referred to the population at the beginning of the period;
- S2 is the vector containing official cross-sectional estimates referred to the population at the end of the period;
- D is the vector containing cross-sectional estimates referred to the beginning of the period for those who died during the period;
- L is the vector containing cross-sectional estimates referred to the beginning of the period for those who left the country during the period;
- C is the vector containing cross-sectional estimates referred to the end of the period for those who became 15 years old during the period;

- E is the vector containing cross-sectional estimates referred to the end of the period for those who entered in the country during the period.

| Total Longitudinal Population (15 and over) | | Labour Status at the end of the period | | | | Deaths | People Leaving the Country (15 and over) | Total cross-sectional Population at the beginning of period (15 and over) |
|---|---|---|---|---|---|---|---|---|
| | | Employed | Unemployed | Inactive | Total | | | |
| Labour Status at the beginning of the period | Employed | ( T ) | | | | ( D ) | ( L ) | ( S1 ) |
| | Unemployed | | | | | | | |
| | Inactive | | | | | | | |
| | Total | | | | | | | |
| Children age 15 | | ( C ) | | | | | | |
| People Entering the Country (15 and over) | | ( E ) | | | | | | |
| Total cross-sectional Population at the end of period (15 and over) | | ( S2 ) | | | | | | |

**Figure 3. Scheme for a "desirable" complete matrix with stocks and gross flows from LFS.**

However, mobility between different municipalities (internal migration) can occur, even within the population still resident in the country at the end of the reference period. The levels of internal mobility may be particularly high in some areas and practically negligible in others.

In most EU countries, the current LFS sample design does not "follow" for re-interview individuals who move out of the household/address/dwelling, as this is not required by Eurostat regulations. Moreover, information on deaths of individuals in the initial sample is rarely available. When producing flows estimates using the LFS, therefore it should be considered whether its longitudinal sample represents the whole initial population or only a part of it.

The Italian LFS does not follow up internal movers. Furthermore, data show that internal mobility in Italy is not a negligible phenomenon and that movers are particularly selected with respected to their demographic characteristics (Istat, 2008). Previous studies on longitudinal data from 1993 to 2003 showed also that the behaviour on the labour market of movers and non-movers is quite different (Albisinni and Discenza, 2002).

Bearing this consideration in mind, it should become evident that using the initial population as the reference population for the LFS longitudinal sample is not correct. If we were to weight the longitudinal sample to the initial population we would be making a very strong assumption, i.e. that *the behaviour of movers within the country from one wave to another is similar to that of non-movers*.

The longitudinal component (sub-sample) of the Italian LFS requires thus the specification of a suitable reference population.

Given that the longitudinal data concerns only individual residing in the same municipality both at the beginning and at the end of the period, the reference population should be defined as the population which is resident in the same municipality for the period of 12 months (or 3 months), thus net of deaths and of internal or international migration (Fig. 4)

Considering the following quantities referring to the a generic "province" *p*, gender *s* and 5-years age groups *e*:

$_1P_{pse}$ is the population at the beginning of the period,

$m_{pse}$ are the deaths occurring during the period,

$n_{pse}$ are the births occurring during the period,

$c_{pse}$ is the internal emigration from province *p* during the period,

$i_{pse}$ is the internal immigration into province *p* during the period,

and given the equation

$$_1P_{pse} - m_{pse} - c_{pse} + n_{pse} + i_{pse} = {}_2P_{pse} \tag{1}$$

the longitudinal population $_lP_{pse}$ is defined as follow

$$_lP_{pse} = {}_1P_{pse} - m_{pse} - c_{pse} \tag{2}$$

In principle there are several possibilities to define the reference population for the longitudinal LFS sample, but the final choice depends on the sample design, on the survey design and on the availability of population totals for weighting (Ceccarelli *et al*., 2012). This also has a direct effect on the transition matrix that can be built.

The strategy followed at ISTAT is to provide flows estimates from LFS (denoted with T in Fig. 3) combining two transition matrices obtained through two distinct methods. The first transition matrix contains stock and flow estimates (obtained from weighted longitudinal micro-data) for the longitudinal population. This component, which represents more than 96% of the total population still resident in the country, provides very accurate estimates with many possible

| Longitudinal Population (15 and over) | | Labour Status at the end of the period | | | | Deaths | People Leaving the Country (15 and over) | Total cross-sectional Population at the beginning of period (15 and over) |
|---|---|---|---|---|---|---|---|---|
| | | Employed | Unemployed | Inactive | Total | | | |
| Labour Status at the beginning of the period | Employed | ( TL ) | | | | ( D ) | ( LM ) | ( S1 ) |
| | Unemployed | | | | | | | |
| | Inactive | | | | | | | |
| | Total | | | | | | | |
| Children age 15 | | ( C ) | | | | | | |
| People Moving Across or Entering the Country (15 and over) | | ( EM ) | | | | | | |
| Total cross-sectional Population at the end of period (15 and over) | | ( S2 ) | | | | | | |

**Figure 4: Scheme for a "actual" complete matrix with stocks and gross flows from Italian LFS**

breakdowns (gender, age groups, NUTS region, level of education, etc.). The second transition matrix contains stock and flow estimates for the mobile population. These estimates are obtained for fewer domains by relying on retrospective questions from the cross-sectional sample at the end of the period. Since they are based on both a smaller sample and retrospective information, the estimates referred to the mobile component, which usually represents about 2% of the total population, are characterized by lesser statistical precision and might be less accurate. Thus, the transition matrices currently produced by ISTAT have the scheme shown in Fig. 4, where the following three components differ from those presented in Fig. 3 insofar as:

- TL is the transition matrix containing stock and flow estimates for the longitudinal population (persons still resident in the same municipality at the end of the period);
- LM is the vector containing cross-sectional estimates referred to the beginning of the period for those who moved out of their municipality of residence (both internal and international emigration) during the period;
- EM is the vector containing cross-sectional estimates referred to the last occasion for those who moved into another municipality (both internal and international immigration) during the period.

Thus, the components of the actual complete matrix are obtained as follow:

- S1 and S2 are the official cross-sectional estimates from Quarterly LFS;
- TL are the longitudinal estimates obtained weighting the longitudinal sample;
- D is obtained applying mortality rates (by sex, age and municipality) to the cross-sectional sample referred to the beginning of the period;
- C is obtained, for individuals aged 15 years old, from the cross-sectional sample referred to the last occasion;
- LM is obtained by difference (LM=S1 – TL –D);
- EM is obtained by difference (EM=S2 – TL –C).

## 5. WEIGHTING PROCEDURE AND TREATMENT OF NON-RESPONSES

### 5.1 LONGITUDINAL NON-RESPONSES AND ELIGIBILITY

The longitudinal component of the LFS is affected, even to a greater extent than the cross-sectional sample, by unit non-response, which may occur in subsequent waves and it is of the following typologies:

- Municipality non-response: some (very small) municipalities are substituted in July at the beginning of a new annual survey cycle and some others may, for different reasons, fail to provide the interviews in subsequent waves;
- Household non-response: all the household members refuse to respond to the questionnaire;
- Individual non-response: some household members refuse to respond to the questionnaire, or they cannot be contacted or they started a new household in the same municipality.

Unit non-response may reduce the size of the longitudinal component, thus increasing the variance of the estimates. It can also produce bias if non-respondents have significantly different labour market features compared to respondents (Lohr, 1999; Schafer, 2000; Särndal and Lundström, 2005).

Given the longitudinal population defined as above, and accounting for non-responses, all the individuals interviewed at the initial quarter can be classified according to the scheme of Fig. 5.

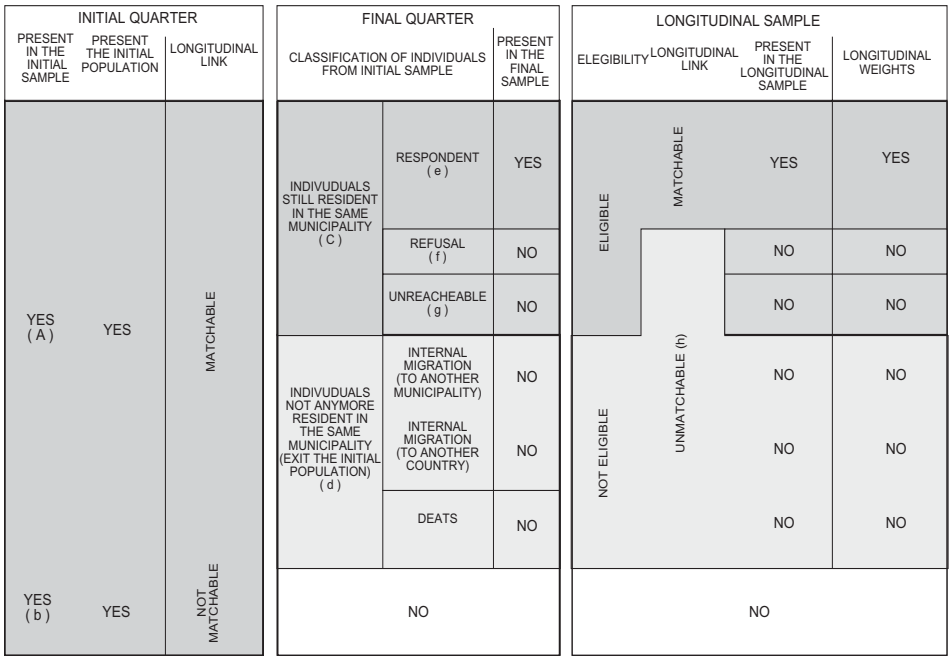


**Figure 5: Classification scheme of individuals from the initial sample and eligibility in the Italian LFS (in presence of longitudinal non-response)**

All the individuals in the sample from the initial quarter (the beginning of the reference period for the longitudinal sample) are divided into two groups:

- *Matchable* (a), if they belong to the rotation groups which are re-interviewed in the final quarter[3];
- *Non-matchable* (b), if they belong to rotation groups which are not wither-interviewed in the final quarter.

Matchable individuals can be further classified into two groups:

- *Eligible* (c): they belong – at least in theory – to the longitudinal population as they are still living in the same municipality, and should therefore be re-interviewed at the subsequent wave. However, some of them are non-respondents in the final quarter either because they refuse to respond (f) or they are unreachable (g), so that they must be considered in a model that aims to treat and reduce the effect of their non-response. These last two groups, in fact, should be

[3] The municipalities that are not in the sample at the end of the period are excluded.

represented by respondents of the final quarter (e) with similar characteristics who have been matched in the longitudinal sample;

- *Non-eligible* (d), they left the initial population during the observed period (deaths and migrations) thus they do not belong to the longitudinal population and must be excluded from a model for treatment of non-response.

When using models/methods for treatment of longitudinal non-response we also face another problematic issue: usually, we do not have enough information to distinguish non-matched individuals who are eligible from not-matched individuals who are not-eligible.

More precisely, we cannot distinguish non-respondents who are eligible (groups f and g in the above scheme) from non-respondents who are non-eligible (d). As immediate consequence, this implies that it is not possible to use standard methods based on logistic regression to compensate for longitudinal non-response in the Italian LFS (Rizzo *et al.*, 1996). The use of these models, in fact, requires that the last subgroup (non-respondents who are eligible) is perfectly identified, in order to adjust the weights of similar matched individuals.

### 5.2 COHERENCE BETWEEN CROSS-SECTIONAL AND LONGITUDINAL ESTIMATES

The last issue concerned with the computation of longitudinal weights is that both cross-sectional and longitudinal estimates referred to the longitudinal population should be produced from the longitudinal sample. The cross-sectional estimates obtained from the longitudinal data should be consistent with the "official" estimates from the (full) cross-sectional sample with reference to the beginning and the end of the observed period.

Fig. 4 shows the transition matrix nested within the quarterly estimates. The difference between the cross-sectional estimates deriving from the cross-sectional sample and those deriving from the longitudinal sample should not be negative. As a matter of fact, these differences refer to the labour force status at the beginning and end of the period, respectively, of the emigrant and immigrant population.

Longitudinal estimates have higher variability than quarterly estimates. It is not possible to ensure their complete consistency with quarterly estimates. However, the weighting strategy applied the longitudinal sample reduces to a considerable extent the risk of obtaining inconsistent results.

## 5.3 WEIGHTING PROCEDURE FOR LONGITUDINAL DATA IN THE ITALIAN LFS

The process to obtain the final longitudinal weights was carried out in three steps (see Appendix) and it made use of two calibration[4] stages (generalised raking procedures (Deville *et al*., 1993):

- the first calibration stage accounted for the differences in the rotation groups and for the bias due to municipality non-response (the all municipality goes out of the sample in subsequent waves). Its application also ensures consistency with quarterly estimates.
- the second stage adjusted for the bias due to individual/household non-response. Furthermore it ensures that the weighted estimates from the longitudinal sample correspond to the figures from the longitudinal population of reference.

### 5.3.1 STEP 1: THE BASE LONGITUDINAL WEIGHTS

In the first step, all the *matchable* individuals are selected; they can be considered like a random sub-sample of the whole cross-sectional sample.

Let $k_i$ denote cross-sectional weight computed for the whole cross-sectional sample at the beginning of the period. Then, for *matchable* individuals, base longitudinal weights $k_i^*$ at different NUTS3 level, are obtained from $k_i$ applying the following correction:

$$k_i^* = \left( k_i \Big/ \sum_{i \in Linkable} k_i \right)_1 P \tag{3}$$

where ${}_1P$ is the cross-sectional population at the beginning of the period.

### 5.3.2 STEP 2: THE INTERMEDIATE LONGITUDINAL WEIGHTS

In order to ensure consistency between longitudinal and cross-sectional "official" estimates, the first calibration stage was applied so that *matchable* individuals at the beginning of the period represented exactly the cross-sectional population of reference of the full cross-sectional sample. Moreover, it made them provide the

---

[4] Calibration involves modifying the original weights in order to simultaneously satisfy several marginal constraints (or to control totals) while minimizing the distance between original and adjusted weights (see Deville and Särndal, 1992; Estevao and Särndal, 2000, 2002; Théberge, 2000)

same cross-sectional "official" estimates for a number of relevant figures cross-classified by sex, region, age group, labour activity status, education, etc. (see Appendix).

Thus, from the base longitudinal $k_i^*$ weight and for all *matchable* individuals, the initial weight $g_i^* = k_i^* \delta_i^*$ was obtained by solving a problem of constrained minimum as follows:

$$\sum_{i \in Linkable} g_{i,rse}^* = \sum_{i=1}^{N_1} w_{i,rse} = {}_1P_{rse} \tag{4}$$

$$\sum_{i \in Linkable} g_i^* X_i = \sum_{i=1}^{N_1} w_i X_i \tag{5}$$

$$\min \left\{ \sum_{i \in Linkable} D\left(k_i^* \delta_i^*, k_i^*\right) \right\} \tag{6}$$

where $w_i$ is the first quarter cross-sectional final weight, ${}_1P_{pse}$ are the known totals for the first quarter cross-sectional population for area *r*, sex *s* and age group *e*, $N_i$ is the first quarter cross-sectional sample size, $X_i$ are auxiliary variables (estimates from full cross-sectional sample at the beginning of the period), D is a logarithmic distance function (Deville and Särndal, 1992).

### 5.3.3 STEP 3: THE FINAL LONGITUDINAL WEIGHTS

The final longitudinal weight $w_i^* = g_i^* \gamma_i^*$, is obtained only for *matched* individuals, applying the second calibration stage. This ensures that the weighted longitudinal estimates of the population are conforming to the known totals of the longitudinal population from official demographic statistics. Thus, the calibration procedure uses the following constraints:

$$\sum_{i \in Linkable} w_{i,rse}^* = {}_1P_{rse} \tag{7}$$

$$\min \left\{ \sum_{i \in Linkable} D\left(g_i^* \delta_i^*, g_i^*\right) \right\} \tag{8}$$

where ${}_1P_{pse}$ is the known longitudinal population.

## 6. CONCLUSIONS

Longitudinal data allow the estimation of labour market transitions, both between the three main labour statuses (employment, unemployment and inactivity) and between more specific occupational conditions (e.g.: from temporary to permanent employment or from full-time to part-time employment). Moreover, also transition and persistence probabilities may be easily derived from the transition matrices.

Therefore, the use of longitudinal data provides an important contribution in terms of assessing the internal dynamics of the labour market that is not deducible from analysis of data cross-section. The longitudinal sample is not a proper panel, consequently it is not representative of the total population. However, it adequately represents the co-present population – which constitutes about 90 % of the initial population - and the derived transition matrix provides very valuable information. A significant improvement would be obtained if we could estimate the transition matrix also with respect to the population who moves between Italian municipalities. The sum of these two matrices would provide a complete picture of the labour market flows for the entire initial population – net of deaths and international emigrations.

The CAPI / CATI technique adopted from 2004 for data collection allows to obtain higher efficiency and accuracy in record linkage, with the use of a unique individual key electronically managed, and correction, through the use of a mechanism for confirmation of the information collected at the first interview to the subsequent interviews, although some issues of lesser importance remain that need to be managed and addressed as set forth above.

The longitudinal population estimation on which to define the structure of the weights allows to produce more realistic estimates hooked to a longitudinal population in the reference period rather than reporting estimates related to the stock of the initial population that, in times of great change particularly due to migration flows, could introduce bias on the estimates. In addition, the weighting procedure is structured in such a way as to take account also of the quarterly estimates already produced in order to make the whole set of information more consistent to the reference frame and comparable with it.

Several problems still remain to be solved such as the estimate for the population moving between municipalities, the estimation of the variances of the gross flow estimates and also to improve the procedures for treatment of non-response (further development in: Kott, 2006; Särndal, 2007), but the point that we discussed in this document is strong enough to allow to produce accurate and meaningful estimates of the dynamic evolution of the labour market.

Longitudinal studies are today an indispensable tool for a more comprehensive and effective labour market analysis. The growing interest in their use is demonstrated

by the many contributions in science – such as those of Mortensen and Pissarides honored with the Nobel Prize for Economics in 2010 – and applications that crop up every day not only in specialised reviews but also on more accessible media.

## APPENDIX

**Step 1 - Calibration constraints on population totals at the beginning of the period known from official demographic statistics.**

| AREA LEVEL | SUB_GROUPS | | Male and Female | Male | Female |
|---|---|---|---|---|---|
| NUTS1 | age_group_14 | | x | x | x |
| NUTS1 | Not Nationals | total | x | x | x |
| NUTS1 | Not Nationals | from EU countries | x | - | - |
| NUTS1 | Not Nationals | from countries Not EU | x | - | - |
| NUTS2 | age_group_7 | | x | x | x |
| NUTS2 | Not Nationals | total | x | x | x |
| NUTS3 | age_group_3 | | x | x | x |

**Step 2 - Calibration constraints on labour status estimated from the full sample at the beginning of the period.**

| AREA LEVEL | SUB_GROUPS | | Male and Female | Male | Female |
|---|---|---|---|---|---|
| NUTS1 | Employed | Total | x | x | x |
| NUTS1 | Employed | Self Employed | x | x | x |
| NUTS1 | Employed | Self Employed - Part Time | x | x | x |
| NUTS1 | Employed | Employees - Permanent | x | x | x |
| NUTS1 | Employed | Employees - Permanent - Part time | x | x | x |
| NUTS1 | Employed | Employees - Temporary | x | x | x |
| NUTS1 | Employed | Employees - Temporary - Part time | x | x | x |
| NUTS1 | Employed | NACE REV1.1 - A - B | x | x | x |
| NUTS1 | Employed | NACE REV1.1 - C - D - E | x | x | x |
| NUTS1 | Employed | NACE REV1.1 - F | x | - | - |
| NUTS1 | Employed | NACE REV1.1 - G - H - I | x | x | x |

**Step 2 (continued) - Calibration constraints on labour status estimated from the full sample at the beginning of the period**

| | | | | | |
|---|---|---|---|---|---|
| NUTS1 | Employed | NACE REV1.1 - J to Q | x | x | x |
| NUTS1 | Employed | Total Not nationals | x | x | x |
| NUTS1 | Employed | Not nationals from EU countries | x | - | - |
| NUTS1 | Employed | Not nationals from Not-EU countries | x | - | - |
| NUTS2 | Employed | Total | x | x | x |
| NUTS2 | Employed | Self Employed | x | - | - |
| NUTS2 | Employed | Employees | x | - | - |
| NUTS2 | Employed | Employees - Permanent | x | - | - |
| NUTS2 | Employed | NACE REV1.1 - A - B | x | x | x |
| NUTS2 | Employed | NACE REV1.1 - C - D - E - F | x | x | x |
| NUTS2 | Employed | NACE REV1.1 - G to Q | x | x | x |
| NUTS1 | Unemployed | Total | x | x | x |
| NUTS1 | Unemployed | Without job experiences | x | x | x |
| NUTS1 | Unemployed | Less than 6 months | x | - | - |
| NUTS1 | Unemployed | from 6 to 11 | x | - | - |
| NUTS1 | Unemployed | 12 months or more | x | x | x |
| NUTS2 | Unemployed | Total | x | x | x |
| NUTS1 | Inactive | Total | x | x | x |
| NUTS1 | Inactive | Search for a job and available in two weeks | x | x | x |
| NUTS1 | Inactive | Search for a job but not available | x<br>x | x | x |
| NUTS1 | Inactive | Do not search for a job but available | x | x | x |
| NUTS1 | Inactive | Do not search for a job and not available | x | x | x |
| NUTS1 | Inactive | Studens | x | x | x |
| NUTS1 | Inactive | Housekeeper | x | - | - |
| NUTS1 | Inactive | Retired | x | x | x |
| NUTS1 | Inactive | Over 65 | x | x | x |
| NUTS2 | Inactive | Total | x | x | x |
| NUTS1 | Isced Level | University Diploma | x | x | x |
| NUTS1 | Isced Level | High School | x | x | x |
| NUTS1 | Isced Level | Secondary School | x | x | x |

**Step 3 - Calibration constraints on the longitudinal population totals known from Official demographic statistics.**

| AREA LEVEL | SUB_GROUPS | | Male and Female | Male | Female |
|---|---|---|---|---|---|
| NUTS1 | age_group_14 | | x | x | x |
| NUTS2 | age_group_5 | | x | x | x |
| NUTS2 | Not Nationals | total | x | x | x |
| NUTS3 | age_group_3 | | x | x | x |

Notes:

age_group_14 = (0-14, 15-19, É, 70-74, 75 and more years)

age_group_7 = (0-14, 15-24, 25-34, 35-44, 45-54, 55-64, 65 and more years)

age_group_5 = (0-14, 15-24, 25-44, 45-64, 65 and more years)

age_group_3 = (0-14, 15-64, 65 and more years)

## References

Albisinni, M., Discenza, A.R. (2002). La mobilità nel mercato del lavoro: principali risultati – aprile 1998-aprile 2002. *Approfondimenti, ISTAT*, Roma.

Ceccarelli, C., Discenza, A.R., Fiori, F. and Lucarelli, C. (2012). Weighting issues in LFS longitudinal data. *Rivista Italiana di Economia Demografia e Statistica*, LXVI (1): 77-84.

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

Deville, J.C., Särndal, C.E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, (88): 1013-1020.

Estevao, V.M. and Särndal, C.E. (2000). A functional form approach to calibration. *Journal of Official Statistics*, (16): 379-399.

Estevao, V.M. and Särndal, C.E. (2002). The ten cases of auxiliary information for calibration in two-phase sampling. *Journal of Official Statistics*, (18): 233-255.

European Commission (1998). Council Regulation (EC) No577/1998 on the organisation of a labour force sample survey in the Community. *Official Journal of the European Union*, Bruxelles: L77/3-L77/7.

Fellegi, I.P. and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, (64): 1183-1210.

Istat (2006). *La rilevazione sulle forze di lavoro: contenuti, metodologie, organizzazione*. Metodi e Norme, ISTAT, Roma.

Istat (2008). *Popolazione e movimento anagrafico dei comuni.* Annuari n. 18, ISTAT, Roma.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, (2): 303-314.

Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32(2): 133-142.

Lepkowski, J.M. (1987), Treatment of wave nonresponse in panel surveys. In: Kasprzyk, D., Duncan,

G., Kalton, G. and Sing, M.P. (Eds.), *Panel Surveys*, Wiley, New York: 348-374.

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

Lynn, P., Buck, N., Burton, J., Jäckle, A. and Laurie, H. (2005). A review of methodological research pertinent to longitudinal survey design and data collection. *Working Papers of the Institute for Social and Economic Research*, paper 2005-29, University of Essex, Colchester.

Mortensen, D.T. and Pissarides, C.A. (1994). Job creation and job destruction in the theory of unemployment. *The Review of Economic Studies*, (61): 397-415.

Paggiaro, A. and Torelli, N. (1999). Una procedura per l'abbinamento di record nella rilevazione trimestrale delle forze di lavoro.'*Working paper n. 15, progetto di ricerca MURST "Lavoro e disoccupazione: questioni di misura e di analisi"*, Cleup, Padova.

Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, (22): 43-53.

Rosati, S. (2004). Longitudinal imputation for the quarterly labour force survey. *Atti della XLII Riunione Scientifica*, Cleup, Bari: 335-338.

Rosati, S. and Boschetto, B. (2013). Longitudinal data editing for the Italian LFS. *Proceedings of Statistics Canada International Symposium 2009.* Gatineau, Quebec. In press.

Särndal, C.E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, (33): 99 -119.

Särndal, C.E. and Lundström, S. (2005). *Estimation in Surveys with Non-response*. Wiley Series in Survey Methodology, Chichester.

Schafer, J.L. (2000). *Analysis of Incomplete Multivariate Data*. Chapmann and Hall/CRC, New York.

Tate, P.F. and Clarke, P.S. (1999). *Methodological Issues in the Production and Analysis of Longitudinal Data from the Labour Force Survey*, GSS Methodology Series, 17. Office for National Statistics, London.

Théberge, A. (2000). Calibration and restricted weights. *Survey Methodology*, (26): 99-107.

Torelli, N. (1998). Integrazione di dati mediante tecniche di abbinamento esatto: sviluppi metodologici e aspetti applicativi. *Atti della XXXIX riunione scientifica SIS. Supplemento alla Rivista di Scritti di Statistica Economica*, Sorrento: 293-304.

Winkler, W.E. (1995). Matching and record linkage. In Cox, B.G., Binder D.A., Chinnappa B.N., Christianson A., Colledge, M.J. and Kott, P.S. (Eds.), *Business Survey Methods*, Wiley, New York: 355-384.