Paper presented to the 8th Workshop on LFS Methodology in Gdansk, May 23–24, 2013

Breaks in the LFS and their treatment: the Italian experience.

(on survey changes in 1992 and 2004, on the transition to new NACE and ISCO, on changes of boundaries in NUTS regions)

A. R. Discenza, F. Gallo, A. Martini, S. Rosati, A. Spizzichino
ISTAT - Italian National Institute of Statistics – Labour Force Division

1. Introduction

The innovation process that a repeated survey has to face during his life has always produced breaks in the time series of certain variables, and will continue to do it in the future.

The Labor Force Survey is conducted in Italy on regular quarterly basis since 1959. Because of its long duration, the wide amount of issues covered, the required international comparability, etc, it has suffered several times of changes that led to the inability to make comparisons with the past, in some cases, even for the most important variables.

In this framework, in the recent years, the LFS Unit in Italy has always striven to back recalculate the past data in order to make them consistent (comparable) as much as possible with current data. As result of this effort, all the major break in the time series have been eliminated thus allowing short and long term analysis for all economic indicators and main aggregates produced by the LFS survey.

The aim of this document is to provide an overview of the changes occurred in the Italian LFS in the last two decades and the solutions adopted in order to be able to disseminate consistent time series and databases. The major changes occurred in the Italian LFS are the following:

- 1992: change of definitions, sample stratification and weighting procedure
- 2004: introduction of the continuous survey
- 2008: transition to the new ISCO and NACE classifications
- 2010: introduction of 3 new NUTS3 and modification of two NUTS2 and NUTS1 regions.

2. The change from “quarterly” to “continuous” in the first quarter 2004

The more complex and greater change in the LFS history occurred in January 2004, when the survey changed from “quarterly” (QLFS) to “continuous” (CLFS). It was completely renewed to fulfill Eurostat Regulations, to improve reliability and quality and to widen the contents for national users. Major changes and improvements were made also to the methodology, organization, data collection strategy, interviewing techniques.

When the project for the new LFS started it was already clear that:

a) the effect of all these innovations, combined with the change in the reference period, and the new definitions of employed and unemployed would have introduced not negligible breaks on the time series (even for the main estimates), and that the new LFS data would have highlighted a seasonal pattern quite different from the old one;

b) it would have made impossible to compare the new data with those disseminated up to the last quarter of 2003, creating serious problems for the long- and short-term analysis, and for the production of seasonal adjusted series;

c) all these changes would make impossible to apply a micro approach for back-recalculation of old time series.

At that time, the Italian NSI planned a “strategy” that would have allowed a back recalculation of the historical series of the main indicators of the labor market, from the 1992Q4 to 2003Q4. The main points of this strategy consists of the following:

- continue to carry on the old QLFS and disseminate their estimates up to the first quarter 2004;

- start to observe the labour market with the new CLFS since the first quarter 2003, and start to produce the new estimates for internal use since then;

Thus, for five consecutive quarters, from 2003Q1 to 2004Q1, the two survey were carried on simultaneously by the LFS Unit, on different samples of households. This overlap period was necessary to fine tune the new survey process, and to collect information about the effect of the change. Thus the 5-double-estimates were used to fit the statistical models needed for back re-calculation of the series referred to the period 1992Q4-2003Q4 (the overlap of at least one year was essential for taking into account the seasonal effects).

The approach adopted was a back recalculation at macro level, with model-based components (Gatto, 2006). Macro level because, as mentioned above, it was not possible to recalculate the "new" aggregates at individual level, due to the absence in the QLFS of the necessary information for the new definitions. Model-based because it was based on econometric techniques of time series analysis. It uses the unobserved component structure of the time series to separately recalculate the cycle-trend (annual and short-term component), the seasonal component and the irregular ones. The rationale behind this approach is that these three components are assumed to be independent, each of them is affected by the change in a different way, and thus can be back-recalculated separately and eventually re-aggregated. The method is based on the hypothesis that the temporal components of the CLFS series are a function of the corresponding components of the QLFS series (which was long enough to be decomposed using traditional time series analysis methods).

Back re-calculation was successfully completed during the first half of 2004, and the results were published at the same time with the first press release of the new continuous LFS (first quarter 2004) and with the seasonal adjusted data of the overall period 1992Q4-2004Q1. Hence, quarterly series of the main aggregates of the labor market were disseminated at NUTS2 level, broken down by sex and five-year age classes. For the employment, time series were produced also for 11 groups of economic activity, professional status, job duration, and full-time part-time. For the unemployment, detailed time series were disseminated for different "duration" for people who search for a job.

3. The survey change in the fourth quarter 1992

In April 2013, a new back recalculation for the period 1977Q1-1992Q3 was added to the current series 1992Q4-2012Q4, obtaining a set of time series covering the last 35 years. These series are perfectly consistent with each other, at all levels of aggregation. Future work is already planned in order to increase the details of information.

3.1. The first time series back re-calculation for the period 1977Q1-1992Q3

For this period, a first time series back recalculation was made already in the past (Gatto, Gennari, Massarelli, 2001) to deal with a break occurred in the 1992Q4, when several methodological aspect changed: sample design, weighting, population totals, definitions, etc.

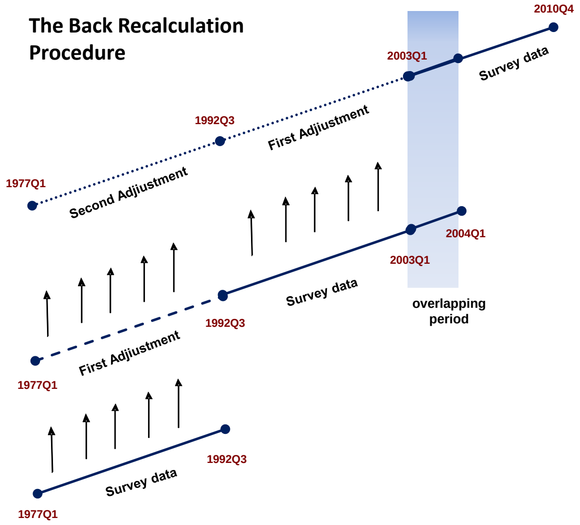
The method, consisted of two steps:

1) In the first step, micro-data of the periods prior to the change were treated to make them as homogeneous as possible to those of later periods (recoding of variables, new imputation method, new weighting procedure), obtaining new time series for the main aggregate by gender and NUTS2, employed in three NACE sectors (agriculture, industry and other activities), employees and self-employed. Moreover, it was possible to recalculate backward the aggregate of the unemployed according to the ILO definition adopted since 1992.
2) In the second step, the series obtained (which continued to have high fluctuations) were "realigned" for consistency with those after the break. The approach to the realignment, operated on the quarterly series, was totally macro in the theoretical framework which realigns separately trend-cycle, the seasonal component and the irregular component of each series (partially similar to the one used for the 2004 break).

This method was recently refined to add more details like broad age groups and more detailed NACE sectors, obtaining a set of time series for the period 1977Q1-1992Q3 consistent with the original time series 1992Q4-2003Q4.

3.2. The second time series back re-calculation for the period 1977Q1-1992Q3

Having obtained a set of realigned time series for the period 1977Q1-2004Q1 (homogeneous in terms of the definitions and methodology used for the period 1992Q4-2004Q1), it was quite straightforward to apply on it the method used in 2004 (explained in Section 2) although with some modifications. In fact, a new step was added such that, using matrix balancing techniques, it has been possible to assure the consistency between new series and past survey data, and preserve the dynamics of the original series.



4. The transition to the new NACE Rev2 classification

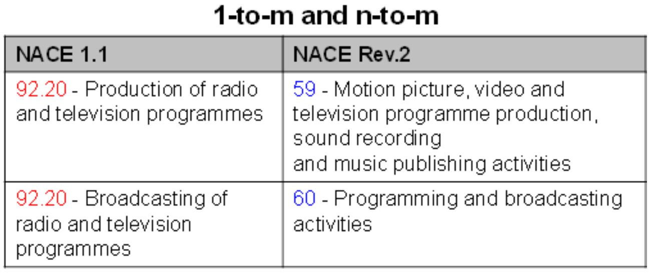
The present Italian Classification of economic activities, called Ateco2007, is derived from NACE Rev.2, which is different in many details from the previous version (NACE Rev. 1.1), although the overall characteristics of NACE remain unchanged. New concepts at the highest level of the classification have been introduced and new detail has been created to reflect different forms of production and emerging new industries. At the highest level of NACE, some sections can be easily compared with the previous version of the classification, while the introduction of some new concepts, e.g. the 'Information section', makes comparison between NACE Rev.2 and its previous version impossible. As a consequence, different types of correspondences between NACE Rev. 1.1 and NACE Rev. 2 can be distinguished: 1-to-1, n-to-1, 1-to-m, n-to-m.

NACE Rev. 2 is to be used for statistics referring to economic activities performed from 1st January 2008 onwards, as established in the NACE Regulation[1]. It is well known that any change of a classification leads to breaks in the time series, which means that statistics are not comparable over time. Nevertheless, methods for adjusting breaks of time series can be used under specific hypothesis. For this purpose, the variable related to economic activity has been coded according to both NACE Rev. 2 and NACE Rev. 1.1 for three years, 2008-2010. The overlap between the two classifications allowed us not only to develop a methodology for adjusting the break of the LFS series, but also to guarantee the continuity with the national account estimates.

The question is: how to automatically get the codes according to both NACE Rev. 2 and NACE Rev. 1.1? The answer to this question is not straightforward, given that the correspondences between the two classifications are not always unambiguous.

During the overlap period the national edition of NACE Rev. 1.1, i.e. Ateco2002, was used as usual classification, but further details were added to the "online vocabulary" in order to establish, for any economic activity, the complete correspondence with NACE Rev. 2. More specifically, units classified in classes associated with 1-to-1 and n-to-1 correspondences can be automatically re-coded, while multiple correspondences were resolved either providing new descriptions for the economic activity, or splitting a

[1] NACE Rev.2 was established by Regulation (EC) No 1893/2006 of the European Parliament and of the Council of 20th December 2006.



1-to-m and n-to-m

NACE 1.1	NACE Rev.2
92.20 - Production of radio and television programmes	59 - Motion picture, video and television programme production, sound recording and music publishing activities
92.20 - Broadcasting of radio and television programmes	60 - Programming and broadcasting activities

single description in more than one associated with different codes of NACE Rev. 2. As a result, a conversion table between the four-digit level of NACE Rev. 1.1 and the two-digit level of NACE Rev. 2 was produced (see the example in the figure).

Such tool allowed us to obtain automatically and simultaneously the pair of codes by choosing the correct description of economic activity during interviews.

The implementation of this strategy takes into account the fact that we use dependent interviewing for interviews after the first. It is worth noting that in this case, all individual records with NACE codes at multiple correspondences were put to blank in the related variable fields of the electronic questionnaire and thus were asked again.

Through this method the code of NACE Rev. 2 was correctly identified. So that, data on the labor market have been collected with the new NACE Rev2- from 2008Q1 and disseminated at national level starting from 2011Q1, when a new specific “online vocabulary” was adopted for 4 digit NACE Rev2. classification.

4.1. Back recalculation of main economic indicator according to NACE Rev2.

This method allowed a more accurate recalculation of a wide set of historical series of main economic indicators in ISTAT.

In the planning stages, it was decided to make use of macro level techniques, for consistency with other series recalculated in the past. The level of disaggregation of economic categories (up to 2 digit) and of the main demographic and economic variables, was chosen taking into account the need of high reliability of the new series. The problems of transition from the old to the new classification affected about 40% of the categories that have multiple correspondences. It was decided to use techniques based on the joint distribution which allow to compute for each category of Ateco 2002 (old national classification) that disaggregates into “n” categories of Ateco 2007 (new national classification) the “weights” of those n categories (incidence over the total).

The application of this method resulted in a comprehensive set of employment data for the period 2004-2007, containing estimates in both Ateco 2002 and Ateco 2007, by gender, NUTS2, professional status and job duration. Moreover, a restricted set of historical series for the period 1977-2010 was recalculated to be used only for seasonal adjustments purposes.

5. The transition to the new ISCO classification

As for the new NACE classification, Eurostat requires that LFS data on occupations have to be sent according to the new Isco08 from 2011. Hence, with the first quarter of 2011, Istat adopted the new national classification of occupations (CP2011), which is completely linked to Isco08.

In accordance with the international classification Isco08, neither the principals nor the first hierarchical level of the classification has been modified. However, even though they maintain the same number and name, the elements they are made up have changed. Nevertheless, the entire framework has been revised in order to provide visibility to the emerging occupations, in particular those involved in the ICT sector.

The adoption of the new Cp2011 not only meant a change in the taxonomy but it also represented a new strategic approach to the occupational coding. In fact, LFS moved from an occupational titles-oriented logic to an approach which uses a detailed description of the activities actually performed by the worker, with 5 hierarchical levels. The on-line coding instrument used by interviewers in their electronic questionnaire have consequently changed: the trigram search within the alphabetical list of occupational titles has been replaced by a search engine which focuses the attention on the work activities actually performed by the respondent.

This change of strategy has been determined by the unreliability of some occupational titles stated by the respondents, often generic or not sufficiently informative. Moreover, a monitoring activity of the coding practices previously adopted by the interviewers, convinced us that the trigram search induced a 'passive' and ‘uncritical’ approach of the interviewer.

The change of the coding strategy together with the substantial innovations introduced by Isco08 and endorsed by Cp2011 made the time-series recalculation quite awkward.

It was not possible to use the same strategy of NACE, but an ex-post recoding was made for all occupations of the interviews collected in 2011, according to the 3-digit of the previous national classification (Cp2001). The double coding was quite straightforward for the 70% of the occupational units, given the 1-to-1 relationship. For the remaining 30%, the situation was much more blurred, especially when the classification effect's go beyond the information stemming from the occupational titles (see for instance the case of managers / supervisors).

To test the quality of this recalculation, longitudinal data were used. In fact, longitudinal information were used for all the individual who did not change jobs between 2010 (when the old ISCO was used) and 2011 (when the new ISCO was used and the old ISCO was recalculated). For this subsample the consistency between the occupational code given in 2010 and the recalculated code in 2011 was checked. Considering the 3-digit codes, the percentage of correct matches is 89.7%. Nevertheless, the percentage increases to 92.7% when the coding refers to the first digit level.

As result, the double coding for the year 2011 gives us the opportunity to go further back in the recalculation process of the historical series of occupation, applying the same methodology used for the economic activity series (explained in section 4).

6. The introduction of three new NUT3 and modification of two NUTS2 and NUTS1 region

The LFS disseminates a number of estimates broken down by the main socio-demographic characteristics, at NUTS 3 level as annual averages. Starting from January 2010, three new NUT3 “province” were created, bringing their number from 107 to 110. Moreover, seven municipalities moved from one NUTS 2 region to another (Marche to Emilia Romagna), and thus from one NUTS 1 to another (Centro to Nord Est).

The main problem to face was that Eurostat do not recognize the new boundaries until January 2012, thus estimates were disseminated with the old boundaries up to 2011Q4, and with the new boundaries from 2012Q1. Moreover, in order to give the possibility to conduct short term analysis , relating to the new territorial unit since from the moment of their creation, LFS Unit has recalculated the estimates produced by the LFS for the four quarters of 2010 and those of 2011. The approach used was that of recoding of micro-data (for all variables related to territorial classification) and subsequent recalculation of the grossing weights taking into account new known totals of population. The method chosen for the calculation of the estimates for new provinces is based on the need to simultaneously achieve the following two objectives:

1. for territorial units (NUTS3, NUTS2 and NUTS1) not affected by changes of boundaries, the quarterly and annual estimates already disseminated should not change at all;
2. original national estimates need not change because the borders of the nation do not change and do not change the reference population as a whole.

For the first objective, the new estimates were obtained by recalculating the weighting factors only for the part of the sample falling in the areas affected by the territorial changes. This was make separately for three distinct spatial domains (e.g. one of this domain refers to the old province of Milano that was split into “Milano” and “Monza”).

The second objective, has been achieved by performing the recalculation of the grossing weights after putting an appropriate set of constraints in the calibration estimator, in addition to the usual quarterly ones, so that

a) the new NUTS3 estimates must reflect the new known total of resident population;
b) new quarterly estimates, referring to each of the three domains as a whole, and regarding the population and the labour market indicators, should remain the same as those already disseminated for the same domains (e.g., the sum of the estimates of the new provinces of Milan and Monza may not differ from the estimate of the old province of Milan).

The constraints necessary to achieve the objective stated in a) are the same as the normal procedure for quarterly estimates (population of the domain by sex and 14 age groups; population of NUTS3 by sex and 5 age groups; non-nationals by sex and citizenship; Number of households of the domain by rotation group; etc).

The additional constraints, summarized below, are the ones that actually allow the achievement of the objective stated in b). These refer to all the indicators disseminated by the LFS, quarterly and annual, at NUTS1, NUTS2 and NUTS3 level, resulting cross-classifying the most relevant labour market and demographic variables (labour status, gender, educational level, age class, nationality, employment status professional status, full-time part-time, sectors of economic activity, occupation, unemployment duration (short, long), previous experience, Neet.

Overall, the calibration process incorporated more than 2 thousands simultaneous constraints and have ensured the achievement of all the objectives 1) and 2). Small differences between new and old estimates were almost exclusively due to rounding effect.

When this recalculation was completed, in March 2013, and at the same time of the 2012Q4 press realease, LFS was able to:

a) update all LFS indicators on datawarehouse I.stat,

b) update all the tables of annual average on the website,

c) provide to users the 8 quarterly the new micro-data files,

d) provide new micro-data files to Eurostat in accordance with the obligations arising from Community changes in the NUTS classification.

Since then, users were allowed to analyze NUTS 3 figures, referring to the new NUTS classification, starting from the year 2010.

References

Gatto R., P. Gennari, and N. Massarelli (2001): "*La ricostruzione e il riallineamento delle serie storiche delle forze di lavoro 1984 - 1992*" Acts of the meeting Occupazione e disoccupazione in Italia: misura e analisi dei comportamenti MURST, Bressanone 15/16 gennaio 2001.

Gatto, R. (2006): "*Series Revision and Seasonal Adjustment of Short Time Series in Presence of a Major Methodological Break*" proceedings of the Conference on Seasonality, Seasonal Adjustment and their implications for Short-Term Analysis and Forecasting, EUROSTAT, Statistical Office of the European Communities, Luxembourg.

Graziani C., Loriga S., A. Martini and A. Spizzichino (2011): *"Il mercato del lavoro in Italia dal 1977: la ricostruzione dei principali indicatori"* AIEL 2011.

Gallo F., P. Scalisi and A. Spizzichino (2012): *"La transizione alla nuova classificazione delle attività economiche: la ricostruzione delle serie storiche e le specificità del settore turismo"*. SIEDS 2012