9th Workshop on Labour Force Survey Methodology
Rome, 15-16 May 2014

The Italian LFS sampling design: recent and future developments

Loredana Di Consiglio, Silvia Loriga, Alessandro Martini, Rita Ranaldi et al.[1]
diconsig@istat.it, siloriga@istat.it, alemartini@istat.it, ranaldi@istat.it

Abstract

The Italian LFS sampling design originally planned in 2002 (to be used in the new continuous LFS started in 2004) has been recently revised, in 2012. Even if the general structure of the sample did not change, some updating and improvements have been introduced jointly with a reduction of the sample dimension due to budget constraints.

Main changes are: the auxiliary information used to distribute the sample (that is the frame of the reference population and the target variables, employed and unemployed people estimated by the LFS) has been updated, taking into account also the new boundaries of the administrative units such as municipalities and provinces; previous experiences on LFS non responses have been considered; improvements on the monthly representativeness of the sample have been introduced; the new PSUs have been selected in a way they overlap as more as possible with the previous PSUs, in order to minimize the impact on the fieldwork (and on the final estimates), but preserving the randomness of their selection; a random rotation of a certain number of PSUs has been designed to be applied every year to maintain the sample updated over time (and to guarantee the substitution of municipalities in which all – or almost all – the households already participated to the LFS).

To conduct this work a task force has been set up in Istat, involving both colleagues from the methodological unit (experts on sampling issues) and colleagues from the LFS units (dedicated to the management of the fieldwork and to the estimation process).

In this paper this revision process of the Italian LFS sample is described. Finally some issues for future development are discussed.

1. Sample design

The Italian LFS sampling design is a two stage sampling design (municipalities are PSUs, households are FSUs) with stratification of PSUs and rotation of FSUs.

In each NUTS 3 domain, PSUs are stratified according to the demographic size. Large municipalities, with population over a given threshold (also called self-representative municipalities - SR), are always included in the sample; smaller municipalities (not self-representative - NSR) are grouped in strata, then one municipality in each stratum is selected with probability proportional to its population. Altogether the strata are 1,097 for a total of 1,111 selected municipalities (in some strata multiple municipalities coexist, the reason is the PSUs rotation which will be explained in paragraph 6), 285 of them are self-representative.

---

[1] The work of revision of the Italian LFS sample has been conducted by a dedicated task force in Istat, involving both colleagues from the methodological unit (experts on sampling issues) and colleagues from the LFS units (dedicated to the management of the fieldwork and to the estimation process): the authors would like to thank all of them: Gianlorenzo Bagatta, Claudio Ceccarelli, Stefano Falorsi, Barbara Boschetto, Michele D'Alò, Filomena De Filippo, Claudia De Vitiis, Lorenzo Di Biagio, Antonio Rinaldo Discenza, Andrea Fasulo, Cinzia Graziani, Francesca Inglese, Rita Lima, Carlo Lucarelli, Alessandro Ortenzi, Daniela Pagliuca, Simona Rosati, Monica Russo, Michele Antonio Salvatore, Andrea Spizzichino, Tiziana Tuoto e Emanuela Vergura.

At the second stage households are randomly selected from the population registers in all the municipalities drawn at the first stage.
The households are rotated according to a 2-(2)-2 rotation scheme. Households are interviewed during two consecutive quarters. After a two-quarters break, they are again interviewed twice in the corresponding two quarters of the following year. As a result, each household is included in four waves of the survey in a period of 15 months.
The size of the theoretical yearly sample is 286.144 households. Every year a new sample of 71.536 sets (blocks) of four households, for a total of 286.144 households, is drawn in order to assure also the substitution of the non-responding units (for each selected household three additional households are chosen as substitute units). The new sample is gradually introduced starting from the first wave of the third quarter. A mixed mode CAPI-CATI is used.

Leaving unchanged this general feature, the sampling design has been recently revised. The new sample has been introduced in 2012Q3; due to the rotation scheme only in 2013Q4 all the four rotation groups have been selected according to the new sampling design.
Several reasons led to the decision to introduce this revision: a) the sample that was in force till 2012 was designed in 2001-2002, considering the target variables (employed and unemployed people) estimated at that time by the quarterly LFS (still not continuous) and the frame information for stratification was referred to 2002; b) it was necessary to update the sample to several changes occurred in the boundaries of the administrative units such as municipalities and provinces; c) it was considered proper to further improve the monthly representativeness of the sample, considering the high relevance of monthly LFS estimates; d) budget constraints made it necessary to reduce the sample size.
To conduct this work a task force has been set up in Istat, involving both colleagues from the methodological unit (experts on sampling issues) and colleagues from the LFS units (dedicated to the management of the fieldwork and to the estimation process).
The new sample has been designed taking into account both methodological and operational constraints: i) the unemployment figures considered as target variables for the evaluation of precision requirements are referred to the pre-crisis period (2004-2007); ii) the information on non responses has been considered when distributing the sample units among the territorial units; iii) the monthly distribution of the sample guarantees that each month (even if composed by 4 or 5 weeks) is representative of the whole national territory; iv) the new selected PSUs have to overlap as more as possible with the previous PSUs in order to minimize the impact on the fieldwork (and on the final estimates); v) a random rotation of a certain number of PSUs has to be applied every year to maintain the sample updated over time (and to guarantee the substitution of municipalities in which all – or almost all – the household already participated to the LFS).
In the following the main issues about the revision of the Italian LFS sample are described; the precision requirements (both European and national) adopted in designing the sample, the distribution of the sample over space (NUTS 3 domains) and time (reference weeks), the strategy to maximize the overlapping between old and new PSUs, the methodology to introduce a random rotation of the PSUs; finally an evaluation of the impact of the introduction of the new sample is shown and some conclusive remarks about future perspectives are outlined.

2. Precision requirements

To design the Italian LFS sample, Eurostat precision requirements (as in Reg. 577/98) have been considered as constraints; additional constraints have been added for national purposes, in particular because of the need to disseminate reliable figures on employment and unemployment in NUTS 3 domains on annual basis. It is worth noting that in recent years the request of information even more disaggregated in territorial domains increased a lot and NUTS 3 LFS estimates produced on annual basis are always observed by local authorities and policy makers, to analyze territorial differences that in Italy are historically well pronounced.

A further constraint on employment at national level has been added in order to balance the national sample taking into account also employment.
The complete set of constraints is shown in Table 1. It is worth explaining that Eurostat precision requirements refer to the relative standard error for a group of unemployed people representing 5% of the working age population (that is aged 15-74 years old), and they are expressed with reference to the estimation of annual averages in NUTS 2 domains and changes between two successive quarters at national level. These constraints have been transformed into equivalent constraints (in terms of final accuracy) for quarterly estimates, these equivalent quarterly precision requirements are shown in the table.
Moreover the differences between Eurostat constraints and Italian constraints for the estimates in NUTS 2 domains and at national level are explained considering that Eurostat constraints refer to a group of unemployed people representing 5% of the working age population, while Italian constraints refer to the actual figures on unemployed people estimated by the LFS.
Another operational constraint is the minimum number of households per PSU which is 48.

Table 1 – Precision requirements of the Italian quarterly LFS sample

	EU constraints	***IT constraint on employed people***	***IT constraint on unemployed people***
NUTS 3 domains			25%
Aosta, Trento, Bolzano (small NUTS 2 domains)	12.44%		16%
Other NUTS 2 domains	12.44%		12%
ITALY	1.83%	0.5%	1.96%
Minimum n. households per PSU			48

When this revision of the Italian LFS sample has been designed, updated auxiliary information has been considered: the updated population frame about the number of residents in each municipality; updated figures on employment and unemployment estimated by the continuous LFS for the period 2004-2007 (the pre-crisis period in which unemployment was lower); moreover the experience about non responses in the LFS has been taken into account in order to define the sample dimension for each NUTS 3 domain (in such a way that the expected sample size is able to guarantee the precision requirements).

3. Distribution of the sample over space

Because of the national precision requirement about unemployment estimates in NUTS 3 domains, the distribution of the sample is not proportional to the demographic size of the domains: in provinces in which unemployment is lower, larger sample size is necessary to guarantee precision requirements, and vice versa. The result is that the sample size is rather variable between NUTS 3 domains. This means also that the goal to produce reliable estimates in NUTS 3 domains implies that the sample deviates from the optimal sample we should have obtained considering just Eurostat NUTS 2 and national precision requirements. In any case, Eurostat constraints are satisfied.
In Table 2 a synthetic picture about the heterogeneity of NUTS 3 domains, in terms of resident population, unemployment rate, quarterly sample size and inclusion probabilities is shown.

Table 2 – Minimum, mean value and maximum of resident population, unemployment rate, quarterly sample size and inclusion probabilities in NUTS 3 domains

	NUTS 3 domains			ITALY
	MIN	MEAN	MAX	
Resident households (N)	24,779	231,841	1,769,720	25,502,535
Unemployment rate % (2004-2007)	2.56	7.29	18.50	7.16
Sample size (n)	192	650	3,408	71,536
Inclusion probabilities (n/N%)	0.12	0.39	3.94	0.28

4. Distribution of the sample over time

The quarterly sample is distributed over time in a way that the sample size is uniformly distributed among the 13 weeks, each stratum is observed at least in 3 weeks per quarter and the monthly representativeness of the sample is guaranteed (when the original sample adopted by the LFS from 2004 was designed also the possibility to impose a weekly representativeness of the sample was studied, but the sample size would have been too large; for this reason a monthly representativeness was chosen).

Some PSUs, the largest, are in the sample all the 13 weeks of the quarter. Other PSUs (among them also some chief towns at NUTS 3 level) are in the sample just 3 weeks per quarter, assigning them reference weeks triplet of weeks in which the distance between them is 4 weeks: for instance a PSU may be observed in reference weeks 1-5-9 or 2-6-10 or 5-9-13 and so on.

In table 3 the distribution of the quarterly sample among 13 reference weeks in NUTS 2 domains is shown.

Respect with the first time the sample of the Italian LFS was designed, that is for the continuous LFS starting from 2004, currently a greater importance has been gained by monthly estimates that in Italy are based on just LFS data. For this reason, leaving unchanged the general structure of the sample, which was originally designed in order to guarantee a monthly representativeness, some efforts have been dedicated to improve the monthly distribution of the sample. In particular the goal was to guarantee that each months is representative, even if composed by 4 or 5 weeks.

The main issue is that months, as they are defined now, are not fixed, but they are composed by a number of weeks (4 or 5) that is variable and depends on the number of Thursdays falling in each solar month. This means that some strata (and then some PSUs) to which the weeks 5 or 9 have been assigned, may fall into different months: week 5 may be included into month 1 or month 2 and week 9 may be included into month 2 or month 3 in the quarter (see Scheme 1).

Scheme 1 – Possible combinations of weeks forming the months in a quarter

Possible combinations	Weeks												
	1	2	3	4	5	6	7	8	9	10	11	12	13
4-4-5													
4-5-4													
5-4-4													

Month 1
Month 2
Month 3

Considering that the chief town in each NUTS 3 domain has usually a different labour market respect to the smaller municipalities, in this revision of the sample, we guaranteed that the chief towns for each NUTS 3 domain which are observed just 3 weeks per quarter, are not to be observed neither in week 5 neither in week 9. In this way it is guaranteed that each month of the quarter, even if composed by 4 or 5 weeks, contains the chief towns for all the NUTS 3 domains. It is worth noting that jointly with the chief town other smaller PSUs in each NUTS domain are always present in each month.

Table 3 - Distribution of the quarterly sample among 13 reference weeks in NUTS 2 domains

NUTS 2	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8	Week 9	Week 10	Week 11	Week 12	Week 13	Quarter
1	396	448	444	404	364	516	396	420	392	464	424	452	412	5532
2	172	180	192	200	172	188	192	172	184	172	188	192	200	2404
3	688	624	676	628	712	616	684	640	712	620	664	632	680	8576
4	352	320	272	276	348	316	288	260	296	332	316	252	280	3908
5	304	264	240	284	288	268	212	304	252	324	232	240	300	3512
6	172	204	176	200	180	208	196	188	212	192	196	188	212	2524
7	180	176	180	208	188	200	160	212	188	200	156	212	192	2452
8	432	440	400	384	408	432	432	360	412	408	436	360	364	5268
9	316	284	336	360	284	252	360	348	284	308	320	360	280	4092
10	112	136	132	144	112	136	136	120	148	104	136	136	108	1660
11	160	160	120	144	176	188	92	148	188	176	124	132	168	1976
12	408	376	380	400	416	412	380	380	408	404	376	432	372	5144
13	88	148	116	120	64	152	112	144	60	108	144	132	88	1476
14	60	96	48	84	60	92	68	84	60	76	100	44	100	972
15	472	392	424	400	420	440	420	384	408	456	416	388	380	5400
16	232	312	296	312	284	260	288	312	332	260	284	276	332	3780
17	164	156	136	180	148	148	120	176	144	172	144	120	180	1988
18	92	116	192	256	92	120	184	260	104	92	184	220	176	2088
19	548	528	496	488	464	508	512	536	456	460	556	468	472	6492
20	172	204	172	156	148	228	192	156	160	196	164	180	164	2292
IT	**5520**	**5564**	**5428**	**5628**	**5328**	**5680**	**5424**	**5604**	**5400**	**5524**	**5560**	**5416**	**5460**	**71536**

5. Maximum overlapping between old and new PSUs

The new sample was designed taking into account also the following operational constraint: the new selected PSUs had to overlap as more as possible with the previous PSUs. The main reason for that is to minimize the impact on the fieldwork. In fact the private firm charged to carry out CAPI interviews on behalf of Istat in the national territory has its network of professional and LFS experienced interviewers selected for the territorial areas according to the municipalities drawn by the previous sampling design. Changing all the PSUs, or the majority of them, would have meant to recruit and to train a lot of new interviewers, with evident effects on the fieldwork and risks on the quality of the final estimates.

The new sample was selected by use of Permanent Random Numbers (PRN) to maximize overlapping with the previous sample.

As the PRN of the previous sampling selection were not available the method suggested by Ernst (2004) was applied to obtain PRNs.

Let $p_{i_h^{t1}}^{t1} = \frac{P_{i_h}^{t1}}{P_{h^{t1}}^{t1}}$ is the probability of selection of PSU i in h^{t1} at time $t1$ (is the proportion of population of stratum h in municipality i), and Z_{i_h} a temporary random number Uniform (0,1).

The retrospective random number are defined as follows:

$X_{j_h^{t1}} = 1 - \left(1 - Z_{j_h}\right)^{p_{j_h^{t1}}^{t1}}$ for the selected PSU,

$X_{i_h^{t1}} = 1 - \left(1 - Z_{j_h}\right)^{p_{j_h^{t1}}^{t1}} \left(1 - Z_{i_h}\right)$ for the other PSUs.

Once the X's are determined, for the new selections same random number are used to select the new sample by means of exponential sampling (see Ohlsson, 1996) by taking the minimum in the strata of

$$\xi_{i_h} = -\frac{\log(-X_{i_h})}{p_{i_h,t2}},\ i_h = 1,\ldots,M_h$$

with $p_{i_h,t2}$ the new proportion of population of PSU *i* in new stratum *h*.
Applying this methodology, 831 municipalities, about 75 percent of the PSUs selected according to the new design, overlapped with the previous PSUs. In this way, during the transition from the old to the new design only few and not relevant adjustments in the fieldwork areas were needed without destabilizing the existent CAPI fieldforce.

6. PSUs rotation

Given the continuity of the LFS, every year a certain number of municipalities (about 10 percent) has to be replaced because all – or almost all – the FSUs in the sampling frame already participated to the LFS. So in order to reduce the statistical burden, in particular for the households living in municipalities with a small number of residents, and to avoid any discretional choice in the PSUs selection, a yearly random rotation of PSUs, belonging to NSR strata, has been introduced to maintain the sample updated over time. So the general strategy about LFS design is to update the stratification of the PSUs about every five years, in order to take into account updated information on their population, and to rotate 10-13 percent of PSUs every year, in order to solve contingent problems in the brief period.
In order to reduce the statistical burden in municipalities with a small number of residents, in the new sampling design PSUs are rotated in sampling strata composed by municipalities whose size is small.
Probabilistic rotation is carried out by applying Permanent Random Number (PRN) and constant shift method (see Brewer et al. 1972, Ohlsson 1995).
Let $X_{i_h}; i_h = 1,\ldots,M_h$ be the PRN associated to the M_h PSU (municipalities) of stratum *h*. New random number are associated to the M_h municipalities of stratum *h* by shifting X_{i_h} $X'_{i_h} = X_{i_h} - c_h; i = 1,\ldots,M_h$.
In strata where one PSU is selected selection is then carried out by means of exponential sampling (see Ohlsson, 1995) by taking the minimum in the strata of

$$\xi_{i_h} = -\frac{\log(-X'_{i_h})}{p_{i_h}},\ i_h = 1,\ldots,M_h$$

where p_{i_h} is the proportion of population of stratum *h* in municipality *i*.
In particular rotation of PSUs is performed at different rate in the strata with the following characteristics:

a) Strata composed by municipalities whose number of resident households is less than 600 then c_h=0.3333
b) Strata composed by municipalities whose minimum household number is less than 575 then c_h=0.125.

According to the methodology previously described, in 2014 143 municipalities have been rotated, about 13 percent of the PSUs that were sample in 2013 (Table 4). The nearly totality of the municipalities with less than 1,000 inhabitants and nearly three out four municipalities between 1,001 and 2,000 inhabitants have been rotated. The impact of the rotation has drastically fallen for municipalities with more 2,000 inhabitants.

Considering that North of Italy, in particular West part, is rich of municipalities with small demographic size, the impact of the PSUs' rotation is higher in this geographic area and lower in Southern Italy.

Table 4 – Number of sample municipalities, number and percentage of rotated municipalities from 2013 to 2014 in IT LFS by demographic size and geographic area

	N. of sample municipalities	N. of rotated municipalities	% of rotated municipalities
Demographic size			
Up to 1,000 inhabitants	55	54	98.2
From 1,001 to 2,000 inhabitants	78	57	73.1
From 2,001 to 10,000 inhabitants	404	30	7.4
From 10,001 to 50,000 inhabitants	426	2	0.5
50,001 inhabitants or more	148	0	0.0
Geographic area			
North-West	308	56	18.2
North-East	233	31	13.3
Centre	178	19	10.7
South	260	24	9.2
Islands	132	13	9.8
Total	**1111**	**143**	**12.9**

7. Accuracy evaluation in the old and the new sample design

The new sampling design has been introduced in the third quarter of 2012 so since the production of the monthly figures referring to July it has been reviewed the estimation procedures to take into account the overlap of the old sampling design and the new one.

The approach applied assumes that the two different sub sample, identified by the year of extraction of household, are independent.

The sub-sample defined according to the old design refers to household extracted before 2012 while those extracted since 2012 define the part that refers to the new design.

For each component so identified it is calculated the initial weight *kj* as the reciprocal of the probability of inclusion of the *j-th* sampling unit.

For the quarterly estimate a correction factor for non-response for each household is also computed, obtained as the inverse of the response rate, in order to balance the distribution of the current sample compared to the theoretical one .

A constraint introduced at this stage of the procedure does not allow the correction factors to exceed the limits defined by the corrector average at NUTS 3 level increased or decreased by 25%. In case this happens the strata are collapsed according to the demographic characteristics in order to bring the correctors within the limits established.

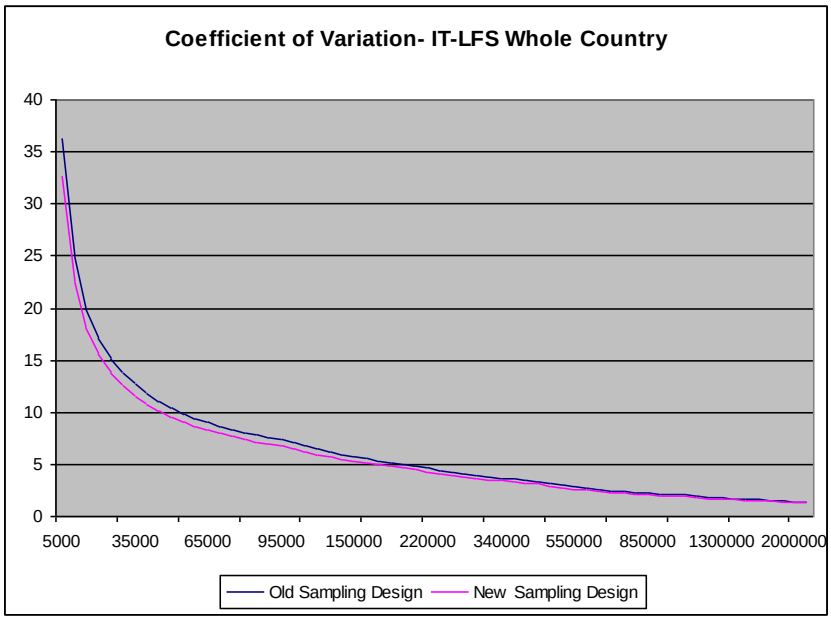
The introduction of new sample has resulted the increase of the strata for which such collapsing was necessary, passing by 3 collapsed strata in the first quarter of 2012 to 16 strata for the last quarter processed, the fourth of the year 2013. Due to the presence of this constraint at provincial level, the variability of intermediate weights, adjusted for non-response, showed just a slight increase.

Before proceeding with the final step of calibration the two sub-samples are combined and the base weights are scaled down according to the theoretical composition of the sample for that quarter.

The adjustment of the weights was carried out taking into account the theoretical composition of the sample firstly since the limited deviation of the actual sample from the theoretical one and secondly for the need of avoiding further complexity in the estimation procedures.

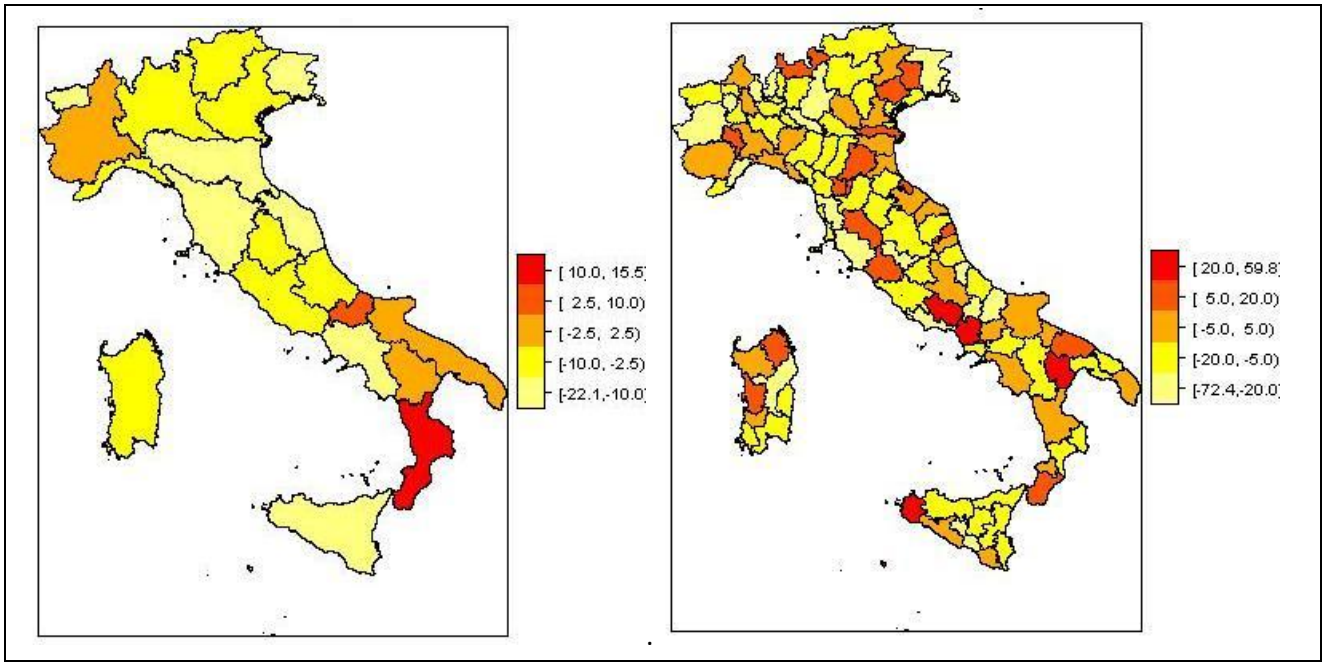
Finally a calibration step is performed so that it is satisfied the condition of equality between the sample estimates and the population totals.
In the results of the calibration procedures, there were no special circumstances: the values of the correction factors remain at levels similar to the past and we do not notice a significant increase in the coefficient of variation of the final weights.
Even in the analysis of variance we took into account the introduction of the new design, according to the same approach. The simultaneous presence of the two designs resulted in an increase in the number of non-aggregatable NSR strata but once established the new design should ensure a better distribution of NSR municipalities.
In the calculation of sampling errors for estimates of annual average, correlation coefficients and the effects of rotation are defined at the regional level and for the total population, therefore, the overlap of the two different designs was not taken into account. This simplification is acceptable because it implies a slight overestimation of the sampling variance since actual overlap of individuals in the sample is lower than that assumed in the definitions of the estimators.
The analysis of the stability of sampling errors is not simple since in this period we observed wide variations in the estimates due to the current economic situation and to the usual seasonal effects as well. In particular the number of unemployed individuals has greatly increased; thereby the corresponding sampling error has been significantly decreased.
In order to have a greater degree of comparability of the results, the analysis was conducted on regression models that fit sampling errors, in order to obtain estimates of sampling errors independently by the observed phenomena, even with an approximate evaluations of the errors.

Graph 1 - Comparison of regression model of sampling errors for the new and old sample design of IT-LFS



Sampling errors at the national level show a slight decline for all levels considered, due to the increased efficiency of the new stratification criterion introduced (Graph 1). Precision level of the estimates, however, is almost stable over the period of overlapping of the two designs.
At regional (NUTS 2), and even at NUTS 3 level, the new sample provides lower sampling errors in almost all the regions, with some exception of regions that were overrepresented in the old sample (Graph 2). In all of them the requirements defined in the planning phase are fulfilled.

Graph 2 - Difference of coefficient of variation between old and new IT-LFS sample design by NUTS 2 and NUTS 3 level



8. Future prospects

In the context of the modernisation of social statistics, waiting for LFS and SILC revision and for the standardisation of variables and modules, Istat has undertaken its process of renewal with the transition to CAPI mode of several PAPI surveys (see for instance SILC and HBS), with the integration of the Trips and Holidays Survey as module into HBS and with the introduction of web in the surveys on PHD graduates and on high school graduates.

The evolution in data collection techniques, that is carrying out several CAPI sample surveys on households, together with the new Population Rolling Census, makes necessary to develop a coordinated approach to develop harmonized sampling designs and to optimize the distribution of the sample over space and time, taking into account the management of the fieldwork. LFS will be involved too. A task force, involving both colleagues from the methodological unit (experts on sampling issues) and colleagues implicated in the different survey processes (Population Census, LFS, SILC, HBS, etc.), has been recently set up in Istat, in order to face in and analyse the issues previously mentioned. It will be a big challenge for Istat.

References

Brewer, K.R.W., Early, L.J. and Joyce, S.F. (1972). "Selecting several samples from a single population", Australian Journal of Statistics, 14, 231-239.

De Vitiis C., Di Consiglio L., Falorsi S. (2005). "Studio del disegno campionario per la nuova rilevazione continua sulle Forze di Lavoro", Contributi ISTAT, Anno 2005 n.6, http://www3.istat.it/dati/pubbsci/contributi/Contr_anno2005.htm.

Lawrence R. Ernst, Yoel Izsak, Steven P. Paben (2004) "Use of Overlap Maximization in the Redesign of the National Compensation Survey", ASA Section on Survey Research Methods

Ohlsson, E. (1995), "Coordination of Samples Using Permanent Random Numbers", In Business Survey Methods, New York: Wiley, 153-169.

Rosén, B. (1997), “On Sampling with Probability Proportional to Size,” Journal of Statistical Planning and Inference, 62, 159-191