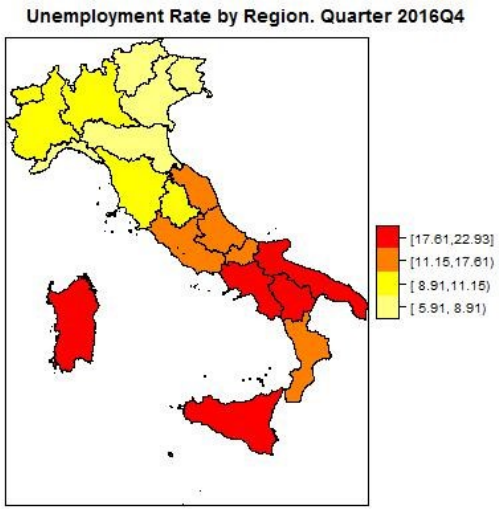
Small Area Estimation of employment and unemployment for Local Labour Market Areas in Italy

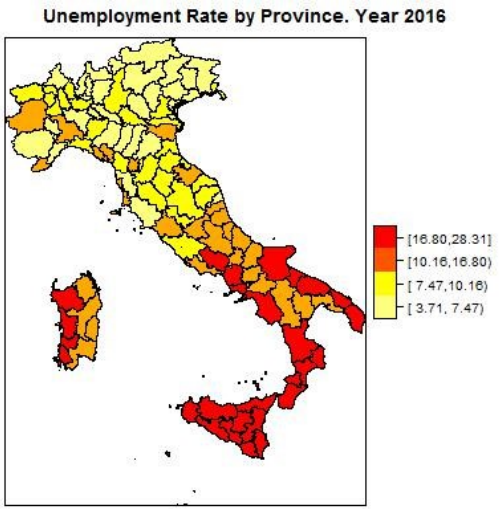
Alessandro Martini alemartini@istat.it, Silvia Loriga siloriga@istat.it

Background

Sample surveys conducted by National Statistical Institutes generally have the purpose of estimating a wide range of parameters (totals and/or averages usually) not only for the entire reference population, but also for specific subpopulations referring to geographic areas or classifications of socio-demographic characteristics. The direct estimates of the parameters, relating to certain subpopulations, are based solely on the data observed on the sample units belonging to these groups.

LFS in Italy (IT-LFS) is designed as a quarterly survey, with rotational pattern 2-2-2, and the sample is uniformly spread across all the weeks such that all territorial domains are represented in each month and in each of the 4 waves (rotational groups). In the cross-sectional perspective it provides monthly estimates, for the whole country, quarterly figures at NUT2 (21 Regions) and yearly figures at NUTS3 level (110 Provinces), as average of the quarterly estimates.





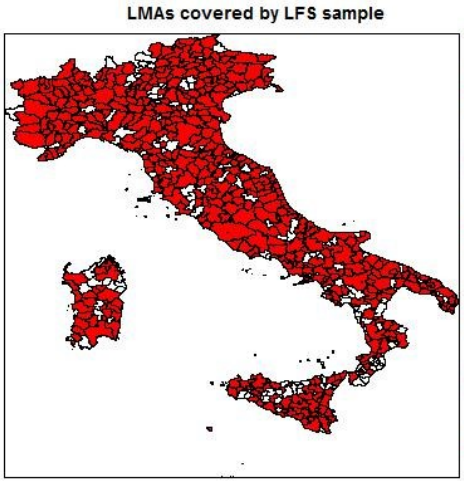
Methodological approach and estimation procedures in IT-LFS take into account the need of providing full consistency of the disseminated figures between the different sets of indicators and with their micro-data.
Even if the sample size of IT-LFS is considerable and the survey is designed to obtain reliable estimates for NUTS3 provinces on annual basis, the sample size it is not adequate to ensure the reliability of the direct estimates for all subpopulations of interest. We use the term *small area* to indicate any subpopulation for which it is not possible to produce direct estimates with a suitable sample accuracy. In the last years there has been an increasing need for local governments (Ministries, Regions, Provinces, Chambers of commerce, Municipalities unions, etc.) of data at sub-regional levels in order to optimize the local planning activities. For these needs many NSIs have in the past given a partial solution increasing sample size in particular subpopulations, for instance in Italy, since 1990, to provide reliable LFS based labour market indicators at sub-regional level, ISTAT increased the basic sample size. In the more recent years Small Area Estimation (SAE) have been developed in order to produce estimates for small domains. These techniques can widely improve data quality in these cases taking into account budget constraints and organizational issues in conducting large scale surveys.

Compared to the direct estimates, *SAE* improve the accuracy level, taking into account the values of the variable of interest observed on sample units of an area, said macro-area, containing the small area in the current and in the previous editions of the survey, allowing to increase the actual sample size to calculate the estimates.

Labour market areas in Italy

Labour market areas (LMAs,"SLL –Sistemi Locali del Lavoro" in Italy) are sub-regional geographical areas where the bulk of the labour force lives and works, and where establishments can find the largest amount of the labour force necessary to occupy the offered jobs. They respond to the need for meaningfully comparable sub-regional labour market areas for the reporting and analysis of statistics. LMAs are defined on a functional basis, the key criterion being the proportion of commuters who cross the LMA boundary on their way to work.
LMAs are developed through an allocation process based on the analysis of commuting patterns. 2011 LMAs are based on commuting data stemming from the 15th Population Census using a new allocation process, an evolution of the previous algorithm: they are 611 distinct areas. Applying this new definition of LMAS we find that for 2015 150 LMAs, 23.7% of the total, are not covered by IT-LFS sample.

Map 1 The 611 Labour Market Areas in Italy



LMAs are not designed to respect administrative boundary constraints: 56 of them (9.2%) cut across NUTS2 regional boundaries and 185 (30.3%) span across different provinces (Nuts3). Voghera and Melfi labour market areas are the only ones cutting across three regions (Nuts2). Milano is the biggest LMA in Italy: it encompasses 3.7 million inhabitants, 174 municipalities belonging to 7 out of the 12 provinces in Lombardy (Nuts2).
In IT-LFS SAE methods have been applied since 1996 to produce LFS-based labour market indicators at Labour Market Areas level, taking into account the definition of LMAs coming from the results of the 1991 population census. According to a design based approach, in the first edition a composite SAE estimator was applied to produce the estimates.
In 2004, in addition to the new definition LMAs (according to the information of the Census 2001), the inferential strategy changed since LFS survey started to be conducted in a continuous way, according to the EU-regulation 577/98. A wide class of SAE methods has been tested, according to the results of EURAREA project, to which Istat participated in those years.
In 2015 a working group[1] has been settled in ISTAT in order to review the estimation methodology according to the new definition of LMA, made available in 2015. The main goals of this working group were:

[1] Waorking group coordinators were Michele D'Alò and Alessandro Martini, other members were: Gaia Bertarelli, Barbara Boschetto, Raffaella Ciocchini, Lorenzo Di Biagio, C.Maria De Gregorio, Dario Ercolani, Stefano Falorsi, Andrea Fasulo, Luisa Franconi, Annelisa Giordano, Alessio Guandalini, Francesca Inglese, Silvia Loriga, Cristiano Marini, Maria Giovanna Ranalli, Antonio Michele Salvatore, Fabrizio Solari.

- to test and implement, starting from the method so far applied and considering the existing scientific literature, a new method to produce annual estimates at LMA level assuring an appropriate level of accuracy, stability and repeatability over time;
- defining a precision evaluation for each estimate at local area;
- to produce estimates consistent with the respective estimates published by Istat in the planned estimation domains, taking into account that the LMAs are unplanned territorial domains and intersect in many cases both the regions that the provinces;
- to take into account the sampling design adopted in the IT-LFS survey, based on the well-known 2-2-2 rotational scheme. This implies a partial overlapping of the sample over time, which is equal to 50% if in two consecutive quarters and in the same quarters of two consecutive years. The estimates al LMA level should take into account the correlation between the observations due to the rotational scheme.

Estimates to be produced at LMA level must refer to: the total population residing in households, the population aged lower than 15 years, labour force, employed and unemployed people, non-labour force, activity rate, employment rate, unemployment rate.

Small Area Methods

Model based small area estimation techniques use explicit modeling for relating unit survey data or area direct estimates to a set of auxiliary information. In the unit level model, individual survey data are required for both target and auxiliary information while at population level totals or mean values of auxiliary variables are needed for each small area. When unit level survey data are not available, an area level mixed model estimator can be implemented. Area level models require strong auxiliary information at area level, which should be available for sampled and non-sampled areas.
For the estimation of labour market indicators at LMA level in Italy unit level models have been applied since 2004. This class of models includes area specific random effects that take into account for between area variations beyond that explained by the set of auxiliary information included in the model.

For the back recalculation of the time series 2004-2015 according to the LMAs defined trough the 2011 Population Census data we have compared the results of different SAE methods:

- Spatial EBLUP (SEBLUP)

  the spatial EBLUP is an estimator based on the unit level model in which the covariance matrix of the random area effects depend on the Euclidean distance among the areas. As a consequence estimates are influenced by the neighborhood. This estimator allows to take under control the bias of the estimates and to obtain more efficient estimates when the spatial correlation between the data is not explained by the auxiliary information.

- Spatial-time EBLUP (STEBLUP)

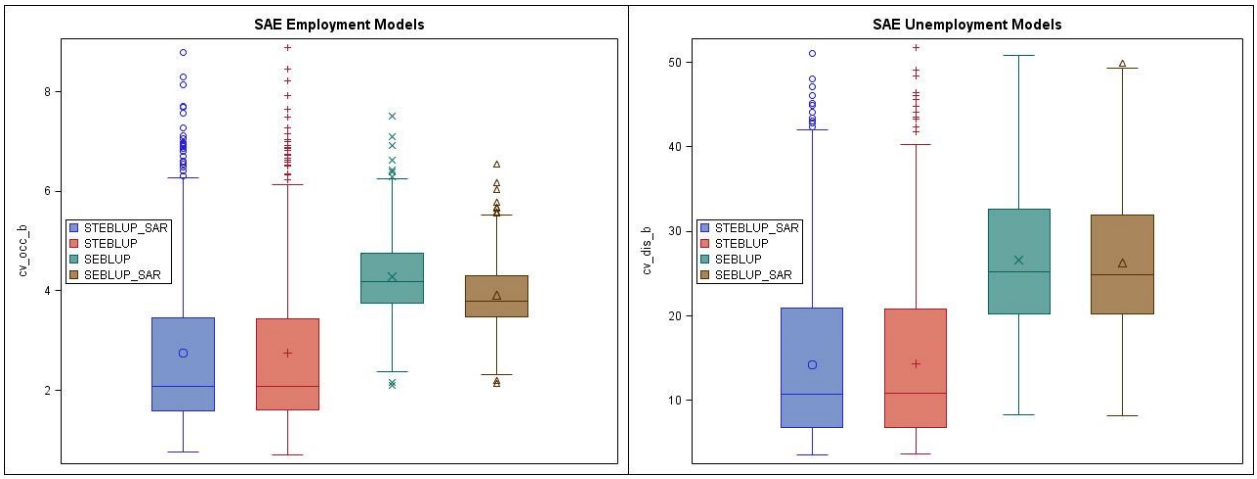
  As in the previous case it considers that the area random effects are spatially correlated. Moreover, it introduces an additional random effect for the time dimension, defined as an autoregressive first order AR (1) process. The consequence is that current edition parameter of interest is linked to the one regarding the previous edition, so the estimates are more influenced both by the neighboring areas (spatial correlation) and by the phenomenon over time (temporal correlation).

For both the two classes of models the covariance matrix of the random area effects has been defined in two different ways, according to:
- the Euclidean distance among the LMAs
- the definition of a neighborhood matrix (SAR).

In the first case the spatial effect depend on the Euclidean distance among the LMAs, in the second case a different distance function takes into account the neighborhood's structure of the LMA.
The box plots in the figure 1 show the Variation Coefficients (CVs) for the different models and specification of the spatial covariance matrix of the random area effects.
The gain of efficiency considering both the space and time correlation is quite clear, CVs are much lower for STEBLUP models than for SEBLUP ones.
Considering that data refer to all quarters since 2004Q1 to 2015Q4 this is quite reasonable: these models can use much more information and are able to provide more precise estimates.
On the other side there are not significant differences considering different matrixes for the correlation structure of the area effects. Considering that the LMAs can slightly vary during the years (municipalities can be born or canceled, with an impact on the LMAs they belong) the Euclidean distance has been chosen, since is a more robust criterion. Otherwise the neighborhood matrix would have been frequently updated, taking into account the binary contiguity criterion, dealing to some estimation issue, since the space correlation structure would not be stable over time.

According to these results STEBLUP estimator and the Euclidean distance as criterion for the covariance matrix of the random area effects were chosen.

Figure 2



Comparison of the reconciliation methods

Methodological approach and estimation procedures in IT-LFS take into account the need of providing consistent estimates for the different indicators. The adopted approach to achieve consistency on both micro data and final figures is based on calibration estimators while benchmarking on macro data is used for seasonal adjusted time series or for small area estimation at LMA level. A benchmarking procedure has been set up in order to take into account contemporaneous constraints according to direct estimates and population totals usually disseminated for planned domains. Through this procedure area model dependent estimates is adjusted considering the direct survey estimate in a group of areas for which the survey estimate is sufficiently accurate.
For IT-LFS, considering yearly figures, NUTS3 level, in Italy called Provinces, are still planned domains, and usually results for these territorial domains are disseminated contemporaneously the press release referring to the last quarter of the year. On a quarterly base planned domains are NUTS2 level (Regions) and, as a consequence of additivity, the whole country.
Variance estimation is an important feature of any small area estimation procedure, in the LMA level MSE estimate a further component has been taken into account the variability, due to the reconciliation of the

small area estimates to the direct ones over a group of areas. However, this procedure inflates the variances of the benchmarked estimates and deals to CVs that can be too high if base SAE estimates are forced to vary much in the reconciliation procedure.

The goal was to find a suitable compromise solution to deal with three main issues:

- define benchmarked estimates consistent with LFS direct estimates at a certain NUTS level, possibly at sub regional level
- do not let final CVs increase too much changing base SAE estimates
- take into account the geographical differences between NUTS classification and LMAs.
- different methods to restore additivity have been compared

Different methods have been applied, considering prorating adjustment, Denton's movement preservation principle and the first one seems to change the base estimates not as much than the others.
According to the first criterion adopted, with which the estimates referring to 2014 were produced, the benchmark was imposed in the case in which aggregations of LMAs coincide with a province or a region, or one their aggregation.
The rule, formally elegant, had two flaws:

- created very uneven areas (Trieste, Val d'Aosta, Sicily, Sardinia, part of Puglia, and a big block in the central area of the country) difficult to interpret;
- few variability at the local level in terms of dynamics of net changes.

Provisional data for 2015 have been defined to be analyzed in the ISTAT Annual Report 2016, and in light of previous results a regional benchmarking was applied (considering as benchmark estimates aggregation of LMAs in which the centroid municipality belongs to a certain region).
Again there was evidence of a trend of regional benchmarking to bring back, very often, the observed heterogeneity at the level of the SLL to the behavior of the reference region. This typically happens in the presence of higher variability of the provincial estimates, probably due even to a quite high sample error.
We therefore sought to define a policy to apply a sub-regional benchmarking, trying different aggregation techniques of LMAs, in order to determine a compromise solution that would provide a framework as consistent as possible for users. The results obtained, refined in several attempts, showed a greater sub-regional heterogeneity and improved both consistency with the pseudo-provincial level estimates and the accuracy of the estimates in some provinces. Some issues anyway arose, trying to apply the benchmarking at province level, in particular regarding the inflation of CVs.
So it has been defined a mixed criterion for benchmarking, combining the two types of benchmarking, at regional and province level according to a distance criterion between initial and final CV. For the unemployment figures of 2016, it is observed that:

- as a consequence of more detailed constraints the average CV increases. LMAs SLL with CV greater than 40% increased from 2 to 48 with the mixed criterion for benchmarking
- a significant effect on the maximum values of CV (43% to about 99% for 2016 figures)
- difficulty in determining unique and general rules to define which criteria to adopt, given that for some cases the different geography SLL vs province can determine uninterpretable or unreliable estimates.

As for the accuracy evaluation of the LMA estimates two additional components of variability were considered while in the previous LMAs estimation methodology were ignored. The first component takes into account the correlation due to the rotational scheme adopted in IT-LFS sampling design, in order to obtain an estimation of the rotation effect the rotation group has been considered, through the addition of a dummy variable for wave groups 2, 3 and 4. The second component deals with reconciliation of the base SAE estimates for LMAs. According to this second component MSE estimation is adjusted and includes the benchmarking effect, adding the squared differences between original and benchmarked estimates (You,Y ,2004).

According to this adjustment final formula for total MSE is:

$$MSE_{Bench} = MSE_{SAE} + (\check{Y}_{SAE-}\check{Y}_{Bench})^2$$

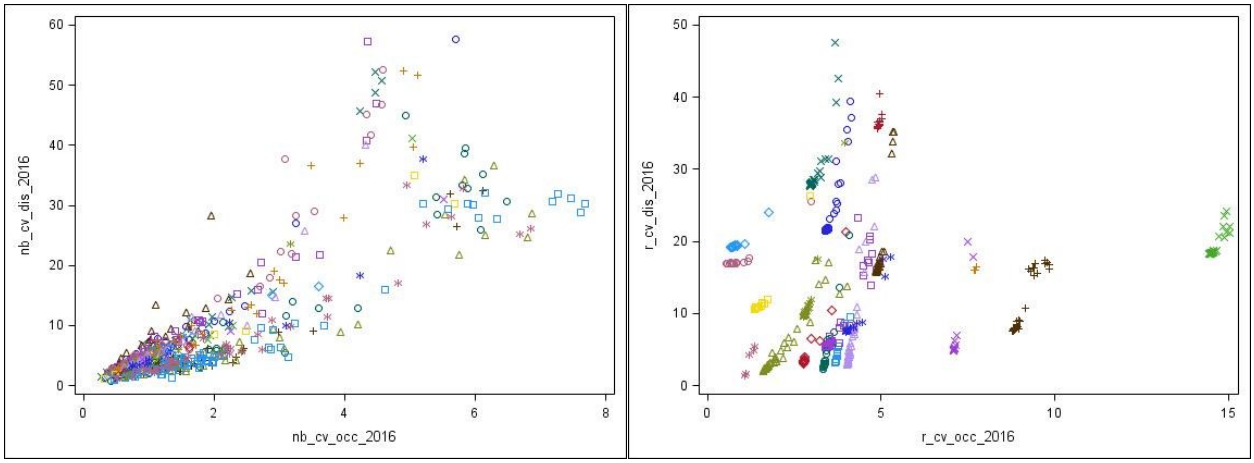
This correction deals to increase CV significantly, in particular in the case benchmarking is applied at sub-regional level.

Table 1 CV statistics for different benchmarking methods. Year 2016

		Regional		Province		Mixed	
		EMP	UNEMP	EMP	UNEMP	EMP	UNEMP
CV	MEAN	5,0	13,7	7,6	19,3	5,4	16
CV	MIN	0,5	1,4	0,3	0,9	0,3	1,0
CV	MAX	15	47,6	44,7	86,4	22,2	86,4

According to these results benchmarking has been applied at regional level, the reconciliation effect of CVs is still relevant, but acceptable. In figure 3 it is clear the effect of benchmarking on CVs, they generally increase, but the results depend greatly by the distance between SAE and direct estimates, aggregated at regional level.

Figure 3 Comparison of Unemployment and Employment CVs pre vs post benchmarking grouped by Region



Preliminary results

Methods defined have been applied in order to back recalculate the series since 2006, since STEBLUP model does not provide estimates for the first two years. They are going to be disseminated in the next future.

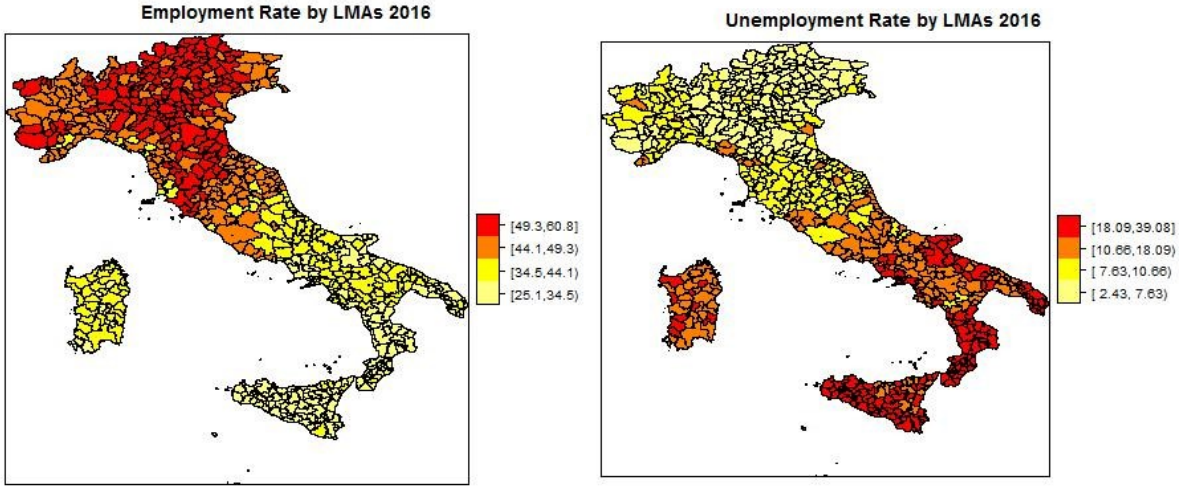
The wide difference in the employment rate[2] values between the center-north of the country and the South provides a first measure of the disparity in employment territorial divide of human resources in the country, this difference for 2016 is more than 15 percentage points. The employment rate, which is 49.7% among residents in the North and 39.9% in the South, while the national average is 43.7%. Even for the unemployment rate differences are considerable: in the North 7.6% of the labor force is in search of employment, in the South is 19.6%, while the national average is 11.7%.

Within these broad areas, however, the differences in the main indicators of the labor market are not very significant, it is therefore necessary, in its analysis of the North-South dualism, analyze regional differences within individual partitions using a finer partition, like that of LMAs.

[2] Main labour market indicators at LMAs level are calculated for the age group 15+

The geographic pattern of the disparities in Italy can be clearly seen in Map 4-5, where Employment and Unemployment rates by LMAs for year 2016 are reported.
LMAs with high employment rate are mainly located in northern part of the country, in particular, the lowest employment rate is in Mondragone and Lentini, both 25.1 %, and the highest is in Brunico, at 60.8 %.
One of the highest dispersion of unemployment rates among European countries is usually observed in Italy, figures at LMA level point this fact out definitely: the LMA of Bagheria has an unemployment rate in 2006 of 39.1 %, more than sixteen times higher than Italy's lowest unemployment rate, which is 2.4 % in the LMA of San Leonardo in Passiria.
For these reasons these data have been defined as one of the parameters to define LMAs eligible as candidates to receive funds provided by a specific law "DM. 181/2016 against crisis in the industrial sector".

Map 4-5 Unemployment and Employment rates by LMAs in Italy, year 2016



Conclusions and future developments

Our goal was to define LMAs labour market estimates consistent with direct estimates of IT-LFS. This has been possible at NUTS2 level taking into account:

- the differences between LMAs and Provinces territorial partition;
- coefficients of variation of benchmarked estimates, too high applying a sub-regional benchmarking criterion.

Moreover benchmarking can also protect the estimates against potential model misspecification and be useful for reducing the over-shrinkage of model based small area estimates.
Auxiliary information or reference indicators at LMAs level availability is still an issue, in the next future data coming from administrative register should increase in our country, in particular for Employment. Model based LMAs labour market indicators estimation could be assessed, improving the detail of disseminated series, according to more relevant age groups and by gender.

References

You, Y., Rao, J.N.K and Dick, P. Benchmarking Hierarchical Bayes Small Area Estimators in the Canadian Census Undercoverage Estimation. Statistics in Transition, 6(5), 631-640 (2004).

Fasulo, M. D'Alò, L. Di Consiglio, S. Falorsi, F. Solari "SMART2: A new web system for Small Area Estimation" ITACOSM2013 conference proceedings http://www.statistica.unimib.it/itacosm13/

M. D'Alò, S. Falorsi, S. Loriga LFS quarterly small area estimation of youth unemployment at provincial level SURWEY Project proceedings http://www.sp.unipg.it/surwey/images/dalo.et.al.sis.ca2014fin.pdf