LFS quarterly small area estimation of youth unemployment at provincial level [*]

Stima trimestrale della disoccupazione giovanile a livello provinciale su dati RFL

Michele D'Aló, Stefano Falorsi, Silvia Loriga

Abstract In this work small area estimation methods are applied to compute quarterly estimates of youth unemployment on Italian provinces. The methods for small areas allow significant efficiency gains compared to the calibration estimator. In addition, they allow us to obtain estimates more in line with Census data. Benchmarking methods have been applied in order to ensure the coherence of the small area estimates with the regular official estimates produced by Istat at higher geographical level of aggregation.

Abstract In questo lavoro sono stati sperimentati alcuni metodi di stima per piccole aree per la stima della disoccupazione giovanile nelle province italiane. Le stime sono state confrontate con quelle dirette, ottenute tramite lo stimatore di calibrazione, e con i dati del Censimento. I metodi per piccole aree mostrano buoni risultati sia in termini di guadagno di efficienza che in base al confronto con i dati censuari. Metodi di benchmarking sono stati applicati al fine di assicurare la coerenza tra stime per piccole aree e le stime ufficiali prodotte dall' Istat a livello di dominio pianificato.

Key words: Mixed models, Spatial and Time Correlation

[*] This work has been supported by project PRIN-2012F42NS8 ”Household wealth and youth unemployment: new survey methods to meet current challenges”

Michele D'Aló
Istat, Via C. Balbo, Rome e-mail: dalo@istat.it.

Stefano Falorsi
Istat, Via C. Balbo, Rome e-mail: stfalors@istat.it

Silvia Loriga
Istat, Viale O. Pacifico, Rome e-mail: siloriga@istat.it

1 Introduction

The aim of this paper is to analyze the possibility to improve the quartely estimation of the labour force status of young people in Italian provinces (NUTS-3) using data from Italian Labour Force Survey (LFS). The target parameter of interest is the number of young unemployed people belonging to the age class 15-24. The main features of the survey and the issues related to the estimation of target parameter are briefly described in Section 2. In Section 3, basic and enhanced small area estimators based on area and unit-level mixed models are considered. The mixed models borrow strength from other areas through the relation of the target variable with a set of available auxiliary information. A specific area random effect is considered in order to account for the heterogeneity among the areas. Enhanced unit level are considered to model the spatial and time auto-correlation among data and in order to take into account the informativeness of the LFS sampling design. Results based on LFS data for the fourth quarter of 2011 are given in Section 4. Conclusions and a focus on the consistency of small area estimates with direct ones, currently produced at the higher level are reported in Section 5.

2 The Labour Force Survey and provincial estimates

The LFS is the main source of information on the Italian labour market that is aimed to produce official estimates of employment, unemployment and inactivity over different territorial domains. Direct estimates are disseminated on monthly, quarterly and annual basis. Only the latter are produced at the finer detailed geographical level, i.e for the provinces. The Italian LFS sample is designed in order to meet precision requirements for the planned domain estimates for a set of given target indicators, as established by the European Council Regulation 577/1998. Additional precision requirements for certain estimates are required only for specific national issues. For instance, for the Italian provinces (NUTS-3) the sample is designed to produce reliable annual estimates of the number of unemployed people (employment and inactivity estimates are always higher and then reliable). Thus quarterly provincial estimates are not produced because in most provinces sampling errors would be too high. Only regional (NUTS-2), macro-area (NUTS-1) and national estimates are computed and disseminated quarterly. In recent years the demand of more frequent and finer information has increased a lot despite of a cut of about 10% of the sample size, due to reduction of the available budget. As a consequence, a study has been performed by ISTAT for computing LFS official estimates by applying Small Area Estimation (SAE) methods. The choice of the target parameter considered in this paper has been done because: this indicator is becoming more and more relevant in the last years, as it is a policy target in the context of the *Youth guarantee* at both EU and national level; the estimates are relatively lower compared with other indicators, hence they are more volatile; the phenomenon is characterized by a strong spatial variability and a pronounced seasonality. The study concerned the use of some small area estimation methods, whose performances in terms of MSE have been compared with the sampling errors estimated for the calibration estimator (Cal.LFS) currently used to produce official LFS figures. Moreover results have been compared with provisional figures from the 2011 Population Census focusing the analysis on North-East provinces. Warnings due to different definitions - such as the reference period, mode and total population - do not influence the comparison among different estimators that use survey data. Besides the gain in terms of MSE, other issues need to be taken into account when model-based small area estimates are computed with the aim to publish official figures. One of the most relevant is the consistency between aggregated SAE estimates and direct estimates currently disseminated for planned domains (in this case, annual estimates in Italian provinces and quarterly estimates in the regions). Moreover, the behavior of the small area estimates should be analyzed also in terms of consistency over time, that is preservation of the seasonal pattern.

3 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. Moreover, direct survey estimates and their corresponding sampling variance need to be available for each sampled area. In this study we experiment estimators based on the unit level mixed model (Battese et al. 1988) and the area level mixed model (Fay and Herriot, 1979). Both models include area specific random effects on the models that account for between area variation beyond that explained by the set of auxiliary variables included in the model. More detail can be found in Rao (2003). Enhanced methods considered are the pseudo-Eblup, the spatial Eblup and the spatial-time Eblup estimators. The pseudo-eblup estimator combines the basic unit linear mixed model with the sampling weights information. With this estimator the informativeness of the design is taken into account, allowing design consistency as the sample size increase and can also be useful to avoid severe bias in the prediction of the variable of interest. Another important advantage is that this estimator allows achieving estimates that are coherent with the direct estimates produced at aggregated level. The spatial EBLUP is an 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. It is possible to introduce other distance functions more suitable for dealing with social surveys or modeling the spatial correlation among data in a nonparametric way. Some examples are given in D'Aló et.al (2012). Finally, the spatial-time Eblup estimator is considered in order to exploit the correlation among data observed in preceding periods, with this estimator a further random effect accounting for the time correlation among data is introduced into the model through an auto-regressive process. That allows gains in efficiency since all the available information can be used for the model fitting. Since different types of models are available model selection of different competitive models has been performed by means of AIC, BIC and conditional AIC and the quality of the estimates have been assessed through bias diagnostics, based on the comparison between direct and model-based estimates (Brown et al. 2001).

4 Application

In this section performances of SAE estimators described in section 3 applied to fourth quarter of 2011 LFS data are shown. The results are comparared with respect to the direct estimates obtained by calibrating on 182 constraints (total population by sex and age classes at the regional and provincial level and for large municipalities, foreign population, number of households). The box plot on the left side of the figure 1 displays the distribution of the CVs% associated to the different estimators, while the box plot on the right side shows the CVs% for the correspondent benchmarked estimates, which allow to compute SAE estimates that reproduce the direct estimates published at regional level. The benchmarking is performed through an expost adjustment and the correspondent MSEs have been augmented by the squared differences between original and benchmarked estimates (You,Y ,2004).
From the figure it is possible to note that small area estimation methods allow to obtain significant efficiency gains especially when the model considers both the spatial and temporal correlation. Finally, the estimates computed through the different estimators have been compared with respect to the correspondent 2011 Census values. This comparison has been carried out by means of Averaged Absolute Relative Error (AARE) and Averaged Squared Error (ASE) and Relative error of the range estimator (RE).

$$AARE\left(\hat{\theta}\right) = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{\hat{\theta}_d}{\theta_d} - 1\right| \tag{1}$$

$$ASE\left(\hat{\theta}\right) = \frac{1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_d - \theta_d\right)^2 \tag{2}$$

$$RE\left(\hat{\theta}\right) = \frac{max(\hat{\theta}_d) - min(\hat{\theta}_d)}{max(\theta_d) - min(\theta_d)} - 1 \tag{3}$$

where for the d-th domain of interest, $d = 1, \cdots, D$, $\hat{\theta}_d$ is the estimate computed with a given estimator, while θ_d is the true parameter of interest for the domain d. All these measures allow a comparative assessment among several competing estimators. The one with the minimum value of these statistic is preferred. The results compared with respect to the provisional data of Census relative to the provinces of

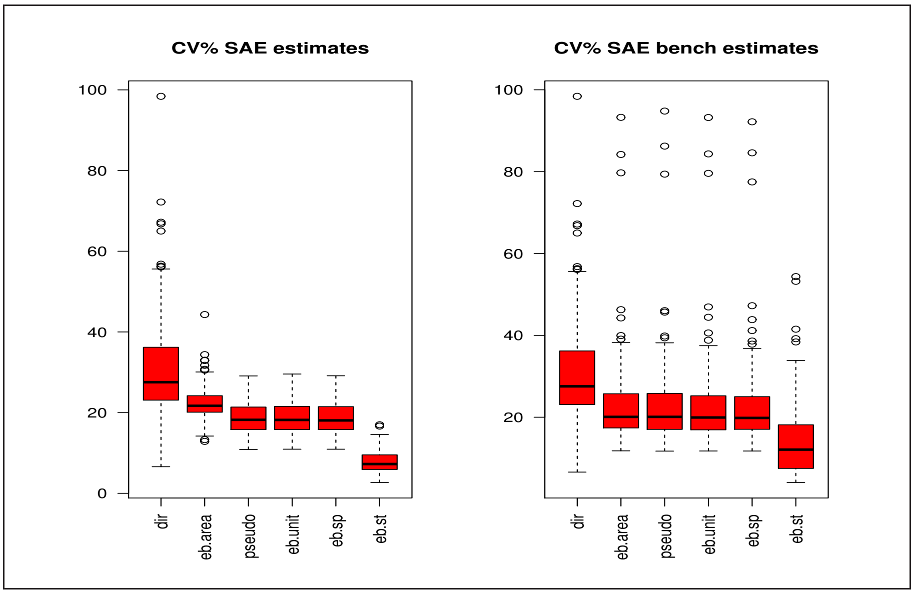

Fig. 1 Distribution CV%

the North-East, are showed in Table 1. S-T.Eblup outperforms the other estimators. AARE and ASE for all SAE estimators (except S-T.Eblup) are very similar. Finally, the bias diagnostics are in line with these results.

Table 1 Estimates comparison respect to the Census data of the provinces Nord-Est Italy

Estimator	AARE	ASE[a]	RE
Cal.LFS	0.390	4694	1.145
Eblup.Area	0.207	1144	0.289
Eblup.unit	0.164	1002	0.238
Pseudo.Eblup	0.169	1104	0.393
S.Eblup	0.164	1001	0.330
S-T.Eblup	0.128	277	-0.045

[a] ASE is divided per 1000

5 Conclutions and future work

The goal of this work is to experiment the use of small area estimation methods in the official statistics production. Quarterly youth unemployment at provincial level has been chosen because of the growing relevance of this indicator and for the increase of the demand of information even more geographically detailed. The results are encouraging and our ongoing work is focusing on an important issue for national statistical offices that is consistency of small area estimates with direct estimates at the higher level usually published and computed through calibration estimator. Our goal is to obtain small area estimates satisfying the benchmarking property. In this way the aggregation of the model-based estimates will be consistent with the direct estimates for a larger area. Benchmarking is also important to protect the estimates against potential model miss-specification and can be useful for reducing the over-shrinkage of model based small area estimates. Even if CVs of benchmarked estimates are higher we have noted that they allow to protect against model misspecification. In order to evaluate how much of the efficiency gain of the spatio-temporal estimator is due to the larger information used to estimate the fixed part of the model or to the time random effect, the standard eblup coefficients have been estimated using the whole information. The results allow to conclude that a great part of the efficiency gain depends on the amount of information used for the estimation of the fixed part of the model. The CVs of the standard eblup estimates became closer to the correspondent CVs of the estimates based on spatial-temporal eblup model. Anyway the latter are still the best. The issue to face with is that the standard eblup estimates computed in this way can be highly biased. This can be also view when the estimates are compared with Census counts. Anyway, after the benchmarking step the bias of these estimates comes back to be acceptable. Therefore we can conclude that behind the importance of computing SAE estimates that are consistent with the less disaggregated direct estimates computed at regional level, benchmarking is also very useful for the robustness of the estimates produced.

References

1. Battese, G.E., Harter, R.M., and Fuller. W.A. An error-components model for prediction of county crop areas using survey and satellite data, Journal of the American Statistical Association, 80, 28-36 (1988)
2. Brown, G., Chambers, R., Heady, P., and Heasman, D. Evaluation of Small Area Estimation Methods An Application to Unemployment Estimates from the UK LFS, Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective, Statistics Canada (2001).
3. D’Alo,́ M., Di Consiglio, L., Falorsi, F, Ranalli, M.G., and Solari, F. Use of spatial information in small area models for unemployment rate estimation at sub-provincial areas in Italy, Journal of the Indian Society of Agricultural Statistics, 66 (1), 43-53 (2012)
4. Fay, R.E., and Herriot, R.A. Estimates of income for small places: An application of James-Stein procedure to census data, Journal of the American Statistical Association, 74, 269-277 (1979)
5. Rao, J.N.K. Small Area Estimation, New York: John Wiley and Sons (2003).
6. You, Y., and Rao, J.N.K. A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights, Canadian Journal of Statistics, 30, 431-439 (2002).
7. 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).