Modèles de Markov latent et modèles de mélanges finis pour l'estimation pour petits domaines

Gaia Bertarelli [1] & M. Giovanna Ranalli [2] & Francesco Bartolucci [3]
& Michele d'Alò [4] & Fabrizio Solari [5]

[1] *Dipartimento di Economia, Università di Perugia, gaia.bertarelli@stat.unipg.it*
[2] *Dipartimento di Scienze Politiche, Università di Perugia, giovanna.ranalli@stat.unipg.it*
[3] *Dipartimento di Economia, Università di Perugia, francesco.bartolucci@stat.unipg.it*
[4] *Dipartimento per l'Integrazione, la Qualità e lo Sviluppo delle Reti di Produzione e Ricerca, ISTAT, dalo@istat.it*
[5] *Dipartimento per l'Integrazione, la Qualità e lo Sviluppo delle Reti di Produzione e Ricerca, ISTAT, solari@istat.it*

Résumé. En Italie, l'Enquête sur la population active (EPA) est menée trimestriellement par l'Institut national de la statistique (ISTAT), pour produire des estimations sur l'état général de la population active aux niveaux national, régional (NUTS2) et de la province (LAU1), respectivement. L'ISTAT diffuse également des estimations annuelles des taux des personnes actives et des chômeurs à un niveau plus fin, donné par les zones du marché de travail local (Labour Market Areas, LMAs). Ces zones sont des agrégations des municipalités et sont définies à chaque recensement en termes de flux quotidien de déplacement dû au travail. En contraste, avec les niveaux NUTS2 et LAU1, les zones LMAs sont des domaines non planifiés et des techniques d'estimation pour les petits domaines sont nécessaires à utiliser. Le caractère continu de l'EPA nous permet d'utiliser une approche longitudinale. Dans ce travail, nous développons une nouvelle méthode d'estimation de petits domaines, en utilisant des modèles de type Latent Markov (LM) dans un cadre Bayésien. Les modèles LM permettent l'analyse des données longitudinales. Dans ces modèles, les caractéristiques d'intérêt et de leur évolution dans le temps sont représentés par un processus latent qui suit une chaîne de Markov, de sorte que les unités statistiques sont autorisées à se déplacer entre les états latents pendant la durée de l'observation. L'estimation est réalisée à l'aide d'un échantillonneur de Gibbs avec l'augmentation des données et le modèle proposé est appliqué pour estimer les taux de chômage trimestriel pour les zones LMAs italiennes en utilisant des données de 2004 à 2014.

Mots-clés. Cadre bayésien, Taux de chômage, Area-level model, Données longitudinales.

1 Motivation and overview

In Italy, the Labour Force Survey (LFS) is conducted quarterly by ISTAT, the National Statistical Institute, to produce estimates of the labour force status of the population at a national, regional (NUTS2) and provincial (LAU1) level. Since 1996 ISTAT produces LFS estimates of employed and unemployed counts at labour market areas (LMAs) level. LMAs 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 are developed through an allocation process based on the analysis of commuting patterns. Since 2011 LMAs are based on commuting data stemming from the 15^{th} Population Census and are now redefined in 611 distinct areas.

Traditional direct estimation requires sufficiently large samples. Unlike NUTS2 and LAU1 areas, LMAs are unplanned domains and direct estimators have overly large sampling errors particularly for areas with small sample sizes. This makes it necessary to "borrow strength" from data on auxiliary variables from other neighbouring areas through appropriate models, leading to indirect or model based estimates. Small Area Estimation (SAE) methods are used in inference for finite populations to obtain estimates of parameters of interest when domain sample sizes are too small to provide adequate precision for direct domain estimators. Statistical models for SAE can be formulated at the individual or area levels. Since 2004, after the redesign of LFS sampling strategy, ISTAT uses an empirical best linear unbiased prediction (EBLUP) estimator based on a unit level linear mixed model with spatially autocorrelated random area effects and where individual covariates, such as sex by age classes, are inserted in the fixed part of the model. As mentioned earlier, in 2011 LMAs have been redefined and this leads to re-thinking the SAE strategy. In particular, in this paper we will consider area level models. Area level data are computationally easier to manage because they are widely smaller in number, in particular with the application at hand of LFS where we have quarterly data available from 2004 to 2014.

The Fay-Herriot model (Fay and Herriot, 1979) is the basic area level SAE model. It combines cross-sectional information at each time for computing the estimate, but does not borrow strength over the past time periods. When longitudinal data is available, the idea is to borrow strength also over time. In the last two decades, several approaches to borrow strength simultaneously in space and in time have been developed for area-level models. We will limit the review here to those developed within a Hierarchical Bayesian (HB) approach to inference. Ghosh et al. (1996) use a time series model to the estimation of median income of four-person families. Datta et al. (1999) apply this model to a longer time series across small areas from the U.S. Current Population Survey using a random walk model. You et al. (2003) apply the same model to unemployment rate estimation for the Canadian Labour Force Survey using shorter time series data and do not consider seasonal adjustments.

In this work we aim at developing a new area level SAE method based on Latent

Markov Models (LMMs, see Bartolucci et al., 2014, for an introduction). In particular, we propose to use this model to estimate quarterly unemployment rates in LMAs from 2004 to 2014 within a HB framework. Area-level SAE models consist of two parts, a sampling model formalizing the assumptions on direct estimators and their relationship with underlying area parameters and a linking model that relates these parameters to area specific auxiliary information. In this work a LMM is used as the linking model and the sampling model is introduced as the highest level of hierarchy. The definition of SAE methods which are able to take into account the non-observable nature of variables of interest is presented in literature only in Fabrizi et al. (2016), but the authors consider just the cross sectional nature of the problem without investigating its time extension. They develop a latent class unit-level model for predicting disability small area counts from survey data.

LMMs, introduced by Wiggins (1973), allow for the analysis of longitudinal data when the response variables measure common characteristics of interest which are not directly observable. The basic LMMs formulation is similar to that of Hidden Markov models for time series data (MacDonald and Zucchini, 1997). In these models the characteristics of interest, and their evolution in time, are represented by a latent process that follows a Markov chain, usually of first order. LMMs model the evolution of the latent characteristic over time and areas are allowed to move between the latent states during the period. LMMs can be seen as an extension of latent class models to longitudinal data. Moreover, they may be seen as an extension of Markov chain models to control for measurement errors.

The proposed SAE method is applied to quarterly data from the LFS from 2004 to 2014. Covariates include sex-by-age population rates, economic characteristics of the small areas, seasonal and temporal trend adjustments. The model is fitted within a Bayesian framework using a Gibbs sampler with augmented data that allows for a more efficient sampling of model parameters. The model finds a good classification in four latent states. Estimates are also compared with those obtained with the classical Fay-Herriot model and with the time series approach considered by Datta et al. (1999) and You et al. (2003). Diagnostic tools are employed to evaluate the goodness of the alternative approaches and a comparison is also made with data coming from the 2011 Population Census.

Bibliography

[1] Bartolucci, F., Farcomeni, A., and Pennoni, F. (2014). Latent Markov Models: a review of a general framework for the analysis of longitudinal data with covariates. *Test*, 23(3):433–465.

[2] Datta, G. S., Lahiri, P., Maiti, T., and Lu, K. L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the US. *Journal of the American Statistical Association*, 94(448):1074–1082.

[3] Fabrizi, E., Montanari, G. E., and Ranalli, M. G. (2016). A hierarchical latent class model for predicting disability small area counts from survey data. *Journal of the Royal Statistical Society: Series A*, 179(1):103–131.

[4] Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277.

[5] Ghosh, M., Nangia, N., and Kim, D. H. (1996). Estimation of median income of four-person families: a bayesian time series approach. *Journal of the American Statistical Association*, 91(436):1423–1431.

[6] MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and other models for discrete-valued time series*, volume 110. CRC Press.

[7] Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation, 2nd Edition.* Wiley Online Library.

[8] Wiggins, L. M. (1973). Panel analysis: Latent probability models for attitude and behavior processes.

[9] You, Y., Rao, J., and Gambino, J. (2003). Model-based unemployment rate estimation for the canadian labour force survey: a hierarchical bayes approach. *Survey Methodology*, 29(1):25–32.