

Small area estimation by hierarchical Bayesian models: some practical and theoretical issues.

Modelli gerarchici Bayesiani per la stima per piccole aree: alcuni problemi teorici e pratici.

Matilde Trevisani, Nicola Torelli

Dipartimento di Scienze Economiche e Statistiche

Università di Trieste

matilde.trevisani@econ.units.it, nicola.torelli@econ.units.it

Riassunto: I modelli gerarchici Bayesiani sono stati recentemente utilizzati nell'ambito dei problemi di stima per piccole aree. E' tuttavia importante sottolineare come la specificazione del modello debba essere adeguata al particolare problema applicativo. Nel presente lavoro si propone un modello Normale-Poisson-logNormale per la stima per piccole aree e, con riferimento all'applicazione sulla stima del numero di disoccupati in sistemi locali del lavoro in Italia, si mostra come tale specificazione sia più convincente sia da un punto di vista teorico che applicativo.

Keywords: Local Labor Markets, disease mapping, unemployment.

1. Introduction

Small area estimation (SAE) concerns statistical techniques aimed to get estimates for small areas (or domains) when survey estimates for these areas are unreliable, (sometimes, cannot even be calculated) because of the limited sample size available. A review of various statistical approaches to small area estimation is in Rao (2003). Recently, various statistical techniques have been proposed that involve the use of models aimed to "borrow strength" over space, over time or using auxiliary information that is supposed to be correlated to the variable of interest. These models can be estimated by using many alternative approaches and one of the most recent proposals has been to use Hierarchical Bayesian Models (HBM). The structure of HBM for small area estimation is somewhat similar to those models used in disease mapping application, and this relationship could be useful for appropriate specification of the models.

In this paper the use of a Normal-Poisson-logNormal model is proposed for estimating the number of unemployed within a small area and we show that this specification can be more appropriate than the standard Normal-Normal model, commonly used in application of HBM to small area problems. A somehow intermediate way between the authors' proposal and the customary specification is provided by the unmatched Normal-logNormal model used by You and Rao (2002).

One of the most relevant information at local level in Italy is the number of unemployed within small areas like municipalities or Local Labor Markets (LLM, i.e. areas including a group of municipalities which share the same labor market conditions, usually much smaller than a province). However, the traditional survey sources of data like the quarterly Italian Labor Force Survey are not well equipped to meet this need. The Labor Force Survey (LFS) is the major source of information on the labor market, but direct LFS estimates are extremely unreliable for small areas like the LLM.

A specification of an appropriate HBM for this situation, which takes into account the nature of the quantities to be estimated, i.e. counts of unemployed, can easily allow for spatial correlation and the use of auxiliary information available at area level from other data sources. In section 2, HBM are introduced and the Normal-Poisson-logNormal is specified. In section 3, the application of this model for estimating unemployment for LLM is considered and some preliminary results are presented.

2. Hierarchical Bayesian models for small area estimation

Let us introduce HBM for SAE problems by revisiting the classical small area model due to Fay and Herriot. Suppose one is interested in estimating the characteristic θ_i , and that the survey estimates $\hat{\theta}_i$ as well as auxiliary data \mathbf{x}_i are known for each area i . Indeed, estimates $\hat{\theta}_i$ are commonly not available for all the small areas (missing data issues will be dealt with in Section 3). The Fay-Herriot model consists of the linking model $\theta_i = \mathbf{x}_i^T \beta + \nu_i$ with $\nu_i \sim N(0, \sigma^2)$ coupled with the sampling model $\hat{\theta}_i = \theta_i + e_i$ with $e_i | \theta_i \sim N(0, \psi_i)$. Sampling variance ψ_i is typically assumed to be known. Classical or empirical Bayes (EB) methods do not require further assumptions whereas HBM require something more. For clarity, hierarchical Bayesian specification of Fay-Herriot model is *fully* written below. The Normal-Normal (N-N) model (as it is referred to in the classical Bayesian terminology) consists of three hierarchical stages,

$$\hat{\theta}_i | \theta_i, \psi_i \sim N(\theta_i, \psi_i), \quad (1)$$

$$\theta_i | \beta, \sigma^2 \sim N(\mathbf{x}_i^T \beta, \sigma^2), \quad (2)$$

$$(\beta, \sigma^2) \sim p(\beta, \sigma^2), \quad (3)$$

with (1) and (2) levels described as above, while (3) involves the further specification of prior distributions for the hyperparameters β and σ^2 . On this regard it is worth noting that, in the absence of substantial prior knowledge, assuming well-defined though extremely vague priors generally ensures a valid estimation.

In HBM approach, inference on the θ_i 's is straightforward and computationally feasible by using standard MCMC-based methods and all the model uncertainty sources are simultaneously accounted for in the estimation process. Moreover, one can draw richer and more realistic models within HBM framework. That is, specification of generalized linear models (GLM) for small area estimation problems is much more feasible than with the alternative methods (Ghosh *et al.* (1998)).

For instance, a common situation arises when the characteristic of interest θ is a total, e.g. the number of unemployed. A Normal-Poisson-logNormal (N-P-IN) model has the following specification:

$$\hat{\theta}_i | \theta_i, \psi_i \sim N(\theta_i, \psi_i), \quad (4)$$

$$\theta_i | \mu_i \sim \text{Poisson}(\mu_i)$$

$$\log(\mu_i) = \log(s\hat{\theta}_i) + \mathbf{x}_i^T \beta + \nu_i \quad \nu_i \sim N(0, \sigma^2), \quad (5)$$

$$(\beta, \sigma^2) \sim p(\beta, \sigma^2). \quad (6)$$

This model is an alternative to the traditional N-N framework, much more convincing when quantities to be estimated are counts (like the number of unemployed). Though this refinement might appear unnecessary when counts $\hat{\theta}_i$ are large enough, however, it is definitely needed when at least some $\hat{\theta}_i$ are ≈ 0 or in general are of small size.

It is also an alternative to the more flexible (GLM) P-IN specification: $\hat{\theta}_i|\theta_i \sim Poisson(\theta_i)$, $log(\theta_i) = log(s\hat{\theta}_i) + \mathbf{x}_i^T\beta + \nu_i + e_i$ with $\nu_i \sim N(0, \sigma^2)$ and $e_i \sim N(0, \psi_i)$. Unlike this, the sampling error e_i is introduced on a separate level, (4), from the linking (non linear) stage (5). Thereby, the unbiasedness assumption $E(e_i|\theta_i)=0$, which fails in (GLM) P-IN model, still holds here. With this regard, the model (4)-(6) may resemble—though it keeps being conceptually and practically different from—the unmatched model suggested by You and Rao (2002) which consists of a sampling model, $\hat{\theta}_i|\theta_i, \psi_i \sim N(\theta_i, \psi_i)$, specified on a different level from the linking model, $log(\theta_i) = log(s\hat{\theta}_i) + \mathbf{x}_i^T\beta + \nu_i$.

Finally, the N-P-IN model has the advantages usually involved by a suitable link transformation of the parameter of interest. That is, after including any prior information about θ_i through an offset (e.g. the *synthetic* estimate $s\hat{\theta}_i$ log-transformed in (5)), inference reduces to estimating the small area-specific relative risk R_i (e.g. of unemployment) with respect to the large-area expected count $s\hat{\theta}_i$ into which the θ_i mean $\mu_i =_S \hat{\theta}_i * R_i$ factorizes. Still, random effects, possibly spatially correlated, are more easily modelled as components of a relative measure (R_i). It is worth noting that such specification is partly borrowed from the disease mapping settings.

3. An application to unemployment estimation in local labour markets

In this section the proposed N-P-IN model is compared to the traditional N-N one in the context of LLM unemployed estimation for the Italian region of Veneto. Specifically, $\hat{\theta}_i$ are the poststratified direct estimates for the LLM whereas the $s\hat{\theta}_i$ are the regional synthetic estimates, both computed by using data from the Italian LFS carried out in January 1999. A naive estimate of the variances ψ_i is considered by assuming data collected by simple random sampling without replacement within each LLM. Since 14 LLM (out of the 51 LLM in Veneto) were not sampled, i.e. comprise municipalities not included as first stage units in the LFS, no direct estimates are available for them. Two strategies are then adopted to face small area missing data: (i) missing $\hat{\theta}_i$ are model-based estimates (as from stage (4)) or, alternatively, (ii) regional estimates $s\hat{\theta}_i$ are imputed to the missing $\hat{\theta}_i$. In either case, an estimate of the mean squared error of $s\hat{\theta}_i$, by using the method proposed by Marker (1995), is the ψ_i associated with these non-sampled LLM. Finally, it is important to remark that aim of this study is merely an illustration of some advantages that model (4)-(6) involves in the simplest conditions. Accordingly, small areas population counts (as they are synthesized by $s\hat{\theta}_i$) are the only auxiliary data considered.

In detail, sampling stage (either (1) or (4)) is specified as $\hat{\theta}_i = \theta_i + e_i$, with $e_i|\theta_i \sim N(0, \psi_i)$; N-N linking stage (2) is specified as $\theta_i =_S \hat{\theta}_i + \alpha + \nu_i$ with $\nu_i \sim N(0, \sigma^2)$; N-P-IN linking stage (5) is specified as $\theta_i \sim Poisson(\mu_i)$ with $log(\mu_i) = log(s\hat{\theta}_i) + \alpha + \nu_i$ and $\nu_i \sim N(0, \sigma^2)$; hyperparameter stage (either (3) or (6)) is specified as $\alpha \sim N(0, 1.0E4)$, $\sigma^{-2} \sim Gamma(.5, .0005)$.

Some results are illustrated in Figure 1 and Table 1.

Figure 1 shows, for each LLM, the coefficient of variation % (CV) of $\hat{\theta}_i$ (squares), $s\hat{\theta}_i$ (circles) and $\hat{\theta}_i^{HB}$ (filled triangles), as they result from N-N or respectively from N-P-IN by

means of strategy (i). N-P-IN (right panel) generally leads to a larger CV reduction than N-N (left and central panels) does. N-N estimates $\hat{\theta}_i^{HB}$ have dramatically large CV for non-sampled LLM: most of LLM with small population size have $CV \gg 100\%$ (in the left panel y -axis range is 0-1000). Instead the correspondent N-P-IN estimates have very low CV (right panel). As for the other LLM areas, the CV reduction is more homogeneous, but still larger for LLM of smaller size, in N-P-IN than is in N-N (right panel versus central panel, both with y -axis 0-120 limited). It is also worth noting that using strategy (ii) to deal with missing data leads to a larger CV reduction. Also in this case, N-P-IN model performs better than N-N does. A further comparison between the performances of the two models can be carried out by two popular Bayesian model checks: the *deviance information criterion* (DIC) (for a description see Spiegelhalter *et al.* (2002)), and the posterior predictive p -value (PP p) defined as $p = P\{\sum_i (y_{i,rep} - \theta_i)^2 / \psi_i > \sum_i (\hat{\theta}_i - \theta_i)^2 / \psi_i | \hat{\theta}\}$ with hypothetical replications $y_{i,rep}$ generated by the posterior predictive distribution under the considered model. According to these criteria (the smaller the DIC and the nearer the PP p to 0.5 the better the model), the proposed model outperforms the N-N one (Table 1).

Figure 1: Estimated CV(%) by a N-N model (left and centre) and a N-P-IN model (right).

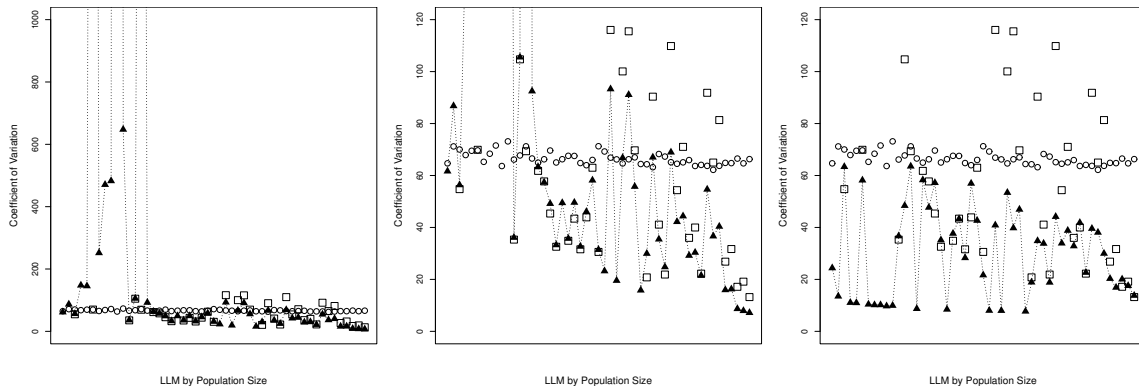


Table 1: Model selection diagnostics.

		\bar{D}	$D(\hat{\theta})$	p_D	DIC	PP p
N-N	(i)	46.1	18.8	27.3	73.5	0.15
	(ii)	44.4	12.5	31.9	76.3	0.11
N-P-IN	(i)	36.8	15.0	21.8	58.6	0.68
	(ii)	41.5	19.2	22.3	63.8	0.67

References

- Ghosh M., Natarajan K., Stroud T.W.F. and Carlin B.P. (1998) Generalized linear models for small-area estimation, *J. Amer. Statist. Assoc.*, 93, 273–282.
- Marker D.A. (1995) *Small Area Estimation. A Bayesian perspective*, Unpublished Dissertation, University of Michigan, Ann Arbor.
- Rao J.N.K. (2003) *Small Area Estimation*, New York: Wiley.
- Spiegelhalter D.K., Best N., P. C.B. and van der Linde A. (2002) Bayesian measures of model complexity and fit (with discussion), *J. Roy. Statist. Soc. B*, 64, 583–639.
- You Y. and Rao J.N.K. (2002) Small area estimation using unmatched sampling and linking models, *Canadian Journal of Statistics*, 30, 3–15.