

Unemployment estimation for Friuli-Venezia Giulia provinces by time series small area models¹

Stima della disoccupazione nelle province del Friuli-Venezia Giulia mediante modelli per piccole aree e serie storiche

Giovanna Menardi, Adriana Monte, Francesco Pauli

Dipartimento di Scienze Economiche e Statistiche²

Università degli Studi di Trieste

email: giovannam@econ.units.it, adrianam@econ.units.it,
francescop@econ.units.it

Riassunto: Tramite l'impiego di tecniche di stima per piccole aree e modelli per serie storiche vengono integrati dati di fonte statistica (Rilevazione Trimestrale sulle Forze Lavoro) e amministrativa (iscritti nelle liste di collocamento) allo scopo di valutare il numero di persone in cerca di occupazione nelle province del Friuli-Venezia Giulia.

Keywords: small area estimation; auxiliary information; labor force; time series models.

1. Introduction

The aim of this work is to estimate the number of job seeker (unemployed in what follows), defined as people who are in a professional condition different from being employed and immediately available to begin a job, in the provinces of Friuli-Venezia Giulia. Two sources of information are available for this purpose. The quarterly labor forces survey (RTFL), carried out by the Italian National Statistical Institute (ISTAT), gives quarterly labor forces estimates at the provincial level which are considered unreliable because the sampling design of the RTFL leads to an insufficient sample size in provinces. In addition, information is available from the regional employment agencies (ARI) on the number of people enrolled in the unemployment lists. The enrollment in the unemployment lists is made on a voluntary basis and does not imply being effectively job seeker as defined above. It is customary to assume, however, that the two variables are correlated. Small area estimation methods are a tool to deal with the situation in which repeated sample survey data are available and offer unreliable but unbiased direct estimate of the quantity of interest. Small area estimators, briefly reviewed in section 2, improve with respect to direct estimators by borrowing strength across neighboring areas, consecutive times and additional information from other sources.

2. Models for Small Area

Small area estimation involves a signal plus noise model

$$Y_{it} = \theta_{it} + e_{it} \tag{1}$$

¹Il lavoro è stato svolto nell'ambito del progetto di ricerca *Metodi statistici per l'integrazione di dati da più fonti*, anno 2002, cofinanziato dal MIUR

²Indirizzo per corrispondenza: P.le Europa 1, 34127, Trieste.

linking the signal θ_{it} , which is the quantity of interest and is unobservable, to its sample counterpart Y_{it} which is observed. Based on (1) and the observed Y_{it} it is possible to estimate the signal θ_{it} by exploiting available information on the sampling errors e_{it} and by making model assumptions which allow to borrow information on θ_{it} from other sources. Direct estimates Y_{it} are usually assumed unbiased, so $E(e_{it}) = 0$, the variance-covariance structure of the errors e_{it} is determined by the sample design.

At the same time, a model links θ_{it} to relevant informations, which may be administrative counts, observations of a related phenomenon, measurements of the quantity of interest in “similar” areas (cross-sectional models), measurements of the quantity of interest on previous occasions (time series) or combinations of the above.

A rather general form of the model is

$$\theta_{it} = \mathbf{x}_{it}^T \beta_{it} + v_i + u_{it} \quad (2)$$

where the variable of interest θ_{it} is related to a set of other variables x_{it} through a linear regression term. The choice of appropriate covariates x clearly depends on the problem at hand. In the context of employment-unemployment estimation administrative counts are often used. In unemployment estimation, in particular, the counts of unemployment benefits claimants is a widely used covariate (Datta *et al.* (1999)).

In the general formulation (2) we included an area-time specific slope β , as in Pfeffermann and Buck (1990); Datta *et al.* (1999) use an area specific slope while Datta *et al.* (2002) assume a homogeneous relationship across areas and estimate a common slope β . These different choices reflect different assumptions on the spatial structure of the phenomenon. Area specific intercepts v_i are generally included in the model.

A time series structure may be included in the model by specifying a time series model for u_{it} . Rao and Yu (1994) use an order one auto regressive model; Datta *et al.* (2002) prefer a random walk model to overcome the difficulties of estimation of the $AR(1)$ model with a correlation coefficient near unity.

A recent general overview of small area models can be found in Pfeffermann (2002), a more detailed methodological description is in Rao (2003).

3. Unemployment estimation for Friuli-Venezia Giulia

The analysis is based on area-level data only, no information is used from individual units. For this reason we do not address the issue of variance estimation for direct estimates, but assume the variances to be known and derive them from ISTAT publications.

Let Y_{it} in (1) be the direct estimate from RTFL for each group $i = 1, \dots, 8$, and time $t = 1, \dots, 28$ corresponding to quarters from the first quarter of year 1993 to the fourth quarter of year 2000. The eight groups are the males and females job seeker in each of the four provinces of Friuli-Venezia Giulia (Gorizia, Pordenone, Trieste e Udine).

The sampling errors e_{it} are assumed independent with known variances (see figure 1(c)). The most recent literature on small area estimation using time series data stresses the importance of taking into account correlation of sampling errors when sampling design is such that units are observed on more than one occasion, as is the case for RTFL. We ignored this issue and assumed the sampling errors e_{it} to be uncorrelated mainly because we lack the data we would need in order to estimate their correlation structure. Moreover, no significant time correlation is estimated on available data. This is likely due to the very small sample sizes involved.

Table 1: Results of model comparison according to deviance information criterion (DIC) and posterior predictive assessment approach (PPAA).

Model	DIC	$\bar{D}$	$\hat{D}$	p_D	PPAA
M1	725.8	639.8	553.7	86	0.589
M2	743.6	629.8	515.9	113.8	0.422
M3	709.5	628.7	547.8	80.9	0.404
M4	707	639.2	565.7	70.8	0.53

The model for θ_{it} is given by

$$\theta_{it} = \beta_i x_{it} + \alpha_i + u_{it} \quad (3)$$

We considered four alternative specifications for α_i , β_i and the errors u_{it} . At first (model M1) we estimated eight separate regression for the eight groups assuming

$$u_{it} \sim IID(t_5(0, \phi_i)); \quad \alpha_i \sim \mathcal{N}(0, 10^6); \quad \beta_i \sim \mathcal{N}(0, 10^6). \quad (4)$$

Within model M1 only the use of administrative counts can reduce the variance of direct estimates. The second model we consider (M2) is a cross-section small area model

$$u_{it} \sim IID(t_5(0, \phi_i)); \quad \alpha_i \sim IIDt_2(\mu_\alpha, \sigma_\alpha); \quad \beta_i \sim IID\mathcal{N}(\mu_\beta, \sigma_\beta). \quad (5)$$

Model M2 ignores the time series structure of the data. The third model (M3) is a cross section - time series model, in fact the errors u_{it} are modeled as an $AR(1)$ process.

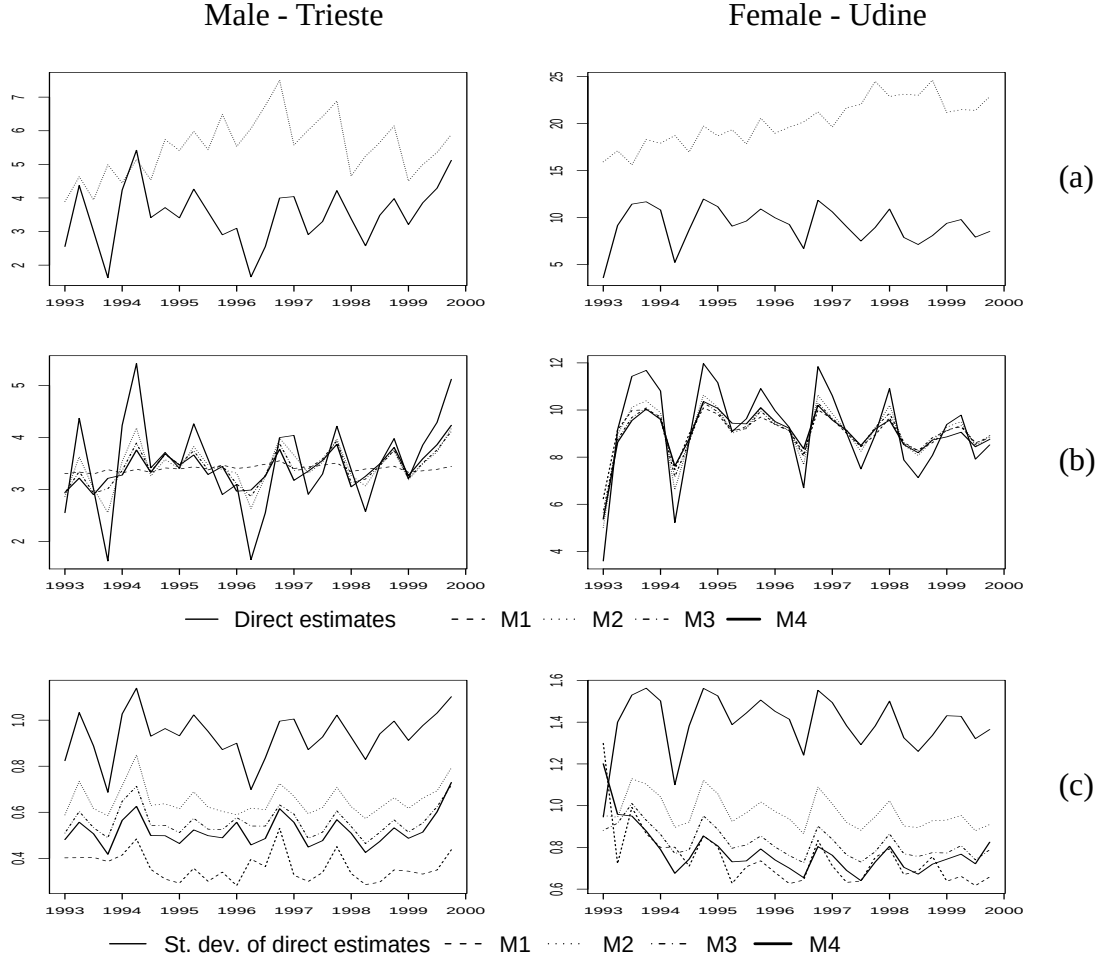
$$u_{it} = \rho_i u_{i,t-1} + \varepsilon_{it}; \quad \varepsilon_{it} \sim t_5(0, \phi). \quad (6)$$

The fourth model (M4) is a modification of model M3 in which we allow for the possibility that in some group the errors follow a random walk rather than an $AR(1)$ process. In practice, a mixture model is used for u_{it} in which $u_{it}|h_{it} = 1 \sim t_5(\rho_i u_{i,t-1}, \sigma_u)$ and $u_{it}|h_{it} = 0 \sim t_5(u_{i,t-1}, \sigma_u)$. To complete the hierarchical bayesian specifications of models M2, M3 and M4, hyperparameters μ_α and μ_β are assumed to be gaussian with mean 0 and high variance, σ_α , σ_β and ϕ_i have independent gamma distributions.

Model comparison (see table 1) is made through posterior predictive assessment approach (PPAA) as in Datta *et al.* (1999) and using deviance information criterion (DIC). In the posterior predictive assessment approach a discrepancy measure, $d(\mathbf{y}; \theta)$, is considered and a set of simulated values of it is compared with the corresponding values calculated on observed data. We set $d(\mathbf{y}; \theta) = \sum_{i=1}^8 (\mathbf{y}_i - \theta_i)^T \Sigma_i^{-1} (\mathbf{y}_i - \theta_i)$, where Σ_i is the diagonal matrix of the variances of direct estimates. We simulate a set of value for y from the posterior predictive distribution $f(\mathbf{y}) = \int f(\mathbf{y}|\theta)\pi(\theta|\mathbf{y}_{obs})d\theta$, calculate $d(\mathbf{y}; \theta)$ on those values and thus obtain the probability $P(d(\mathbf{y}|\theta) \geq d(\mathbf{y}_{obs}|\theta)|\mathbf{y}_{obs})$. If the model is adequate this probability is near 0.5; a value near 0 or 1 suggests lack of fit of the model.

Model M4 turns out to be the most parsimonious in terms of equivalent number of parameters (p_D) and the optimal choice according to DIC, the choice of M4 is also suggested by the PPAA. Posterior means and standard deviations for θ_{it} are shown in figure 1 for all models for two out of eight groups. We note that model M4 does not always offer the highest reduction of variance with respect to variance of direct estimates: in some cases model M1 performs better. However, model choice should not be driven by the degree of variance reduction since a lower variance may be obtained by increasing the bias.

Figure 1: (a): Direct estimates of unemployment (solid line) and administrative counts of job seekers (dotted lines), both in thousands; (b): direct estimates and posterior means of θ_{it} under models M1 to M4 (see section 3); (c): standard deviations of direct estimates and posterior standard deviations for θ_{it} under models M1 to M4 (see section 3).



References

- Datta G.S., Lahiri P. and Maiti T. (2002) Empirical bayes estimation of median income of four-person families by state using time series and cross-sectional data, *Journal of Statistical Planning and Inference*, 102, 83–97.
- Datta G.S., Lahiri P., Maiti T. and Lu K.L. (1999) Hierarchical bayes estimation of unemployment rates for the states of the U.S., *JASA*, 94, 1074–1082.
- Pfeffermann D. (2002) Small area estimation—new developments and directions, *International Statistical Review*, 70, 125–143.
- Pfeffermann D. and Buck L. (1990) Robust small-area estimation combining time series and cross-sectional data, *Survey Methodology*, 16, 217–237.
- Rao J.N.K. (2003) *Small Area Estimation*, Wiley.
- Rao J.N.K. and Yu M. (1994) Small area estimation by combining time series and cross-sectional data, *Canadian Journal of Statistics*, 22, 511–528.