

Measuring the Uncertainty of Small Area Estimators Based On NEF-QVF Superpopulation Models

*Una misura d'incertezza per stimatori per piccole aree
basati su modelli di superpopolazione NEF-QVF*

Michele Crescenzi¹

Dipartimento di Statistica, Prob. e Stat. Appl.
Università di Roma "La Sapienza"
michele.crescenzi@uniroma1.it

Malay Ghosh

Department of Statistics
University of Florida
ghoshm@stat.ufl.edu

Riassunto: In questo lavoro viene considerata una misura d'incertezza per gli stimatori per piccole aree proposti da Sarkar e Ghosh (1998) nel caso di superpopolazione sottostante appartenente alla famiglia naturale esponenziale (NEF) con funzioni di varianza quadratiche (QVF). La procedura di Sarkar-Ghosh risulta interessante perché indica un approccio unificato di Bayes empirico (EB) alla stima per piccole aree in modelli NEF-QVF. Questa famiglia caratterizzata da Morris (1982) include, tra gli altri, il caso binomiale, di Poisson e normale. Dal momento che Sarkar e Ghosh (1998) non forniscono misure di precisione associate alle loro stime, in questo lavoro si ottiene, per queste, un'approssimazione del secondo ordine dell'errore quadratico medio incondizionato di predizione (UMSEP).

Keywords: natural exponential family with quadratic variance functions (NEF-QVF), empirical Bayes (EB) estimation, generalized estimating equations (GEE), Delta method.

1. Introduction

In recent years, small area estimation problems have received growing attention in the statistics literature. This interest is due to the demand of reliable small area or small domain estimates from both the statistical bureaus and many private agencies. In fact, for economic reasons, it is often necessary to provide estimates not only for the total population but also at a much lower level of aggregation from sample survey data targeted to achieve high accuracy at the population level. In such cases direct estimators may not have an acceptable precision because of the smallness of sample sizes of these domains. Thus, in order to obtain reliable small area estimates, it is necessary to "borrow strength" from related areas. This can be done by introducing a model that links the small domains together using some auxiliary information. For a review of the methods proposed in this topic, see Ghosh and Rao (1994).

In this paper we propose a new measure of precision for the empirical Bayes (EB) predictors proposed by Sarkar and Ghosh (1998) to estimate small area means when the underlying superpopulation belongs to the natural exponential family (NEF) with quadratic variance functions (QVF). This family includes, among the others, the binomial and the Poisson distributions, and thus provides solutions when one encounters binary or count survey response variables. Applied statisticians may find this work very useful since most

¹L'autore desidera ringraziare il DSPSA dell'Università di Roma "La Sapienza", per aver fornito i finanziamenti necessari per visitare il Prof. Malay Ghosh presso la University of Florida.

of the present small area estimation literature still concerns normal models. Moreover, the EB analysis avoids the Markov chain Monte-Carlo (MCMC) numerical integration techniques as required by the hierarchical Bayesian (HB) methodology, and so it is easier to implement. We underscore the fact that the EB approach is essentially frequentist because it involves integration over the sample space. Here, after recalling in Section 2 the Sarkar-Ghosh estimation approach, in Section 3 we compute a second-order approximation for a frequentist uncertainty measure: the unconditional mean squared error of prediction.

2. The superpopulation model and the EB small area estimators

In a finite population sampling, suppose Y_{ji} is the characteristic of interest for the i -th unit within the j -th area of the population ($i = 1, \dots, N_j; j = 1, \dots, k$). Our problem consists of predicting the small area population means $\gamma_j = \sum_{i=1}^{N_j} Y_{ji}/N_j$. Let n_j and $f_j = (N_j - n_j)/N_j$ be the sample size and the correction term for finite population of the j -th area. The sample means $\bar{y}_j = \sum_{i=1}^{n_j} y_{ji}/n_j$ are assumed to be conditionally independent ($j = 1, \dots, k$) with pdf's

$$\bar{y}_j|\theta_j \stackrel{ind}{\sim} \exp\{n_j\phi_j[\bar{y}_j\theta_j - b(\theta_j)]\}c(\bar{y}_j, n_j\phi_j). \quad (1)$$

Following McCullagh and Nelder (1989), $E(\bar{y}_j|\theta_j) = b'(\theta_j) = \mu_j$ and $Var(\bar{y}_j|\theta_j) = b''(\theta_j)/(n_j\phi_j) = V(\mu_j)/(n_j\phi_j)$, where $V(\mu_j) = v_0 + v_1\mu_j + v_2\mu_j^2$, with v_0, v_1, v_2 not all zeroes, is the quadratic variance function² and $\phi_j(> 0)$ is supposed to be known for all j .

For the random area effect θ_j we consider the conjugate prior

$$\theta_j \stackrel{ind}{\sim} \exp\{\lambda[m_j\theta_j - b(\theta_j)]\}K(\lambda, m_j), \quad (2)$$

so that the prior mean for the j -th area is $E(\mu_j) = m_j$, while the prior variance is $Var(\mu_j) = V(m_j)/(\lambda - v_2)$, provided $\lambda > v_2$. Next, the posterior distribution of θ_j is given by

$$\pi(\theta_j|\bar{y}_j) \propto \exp\{\theta_j[n_j\phi_j\bar{y}_j + \lambda m_j] - [n_j\phi_j + \lambda]b(\theta_j)\}. \quad (3)$$

It is easy to show that the posterior mean of μ_j is $E(\mu_j|\bar{y}_j) = (1 - B_j)\bar{y}_j + B_j m_j$ with $B_j = \lambda/(\lambda + n_j\phi_j)$.

Following Ghosh and Meeden (1986), the Bayes estimator of γ_j , under squared error loss, is

$$e_j^B = \bar{y}_j - f_j B_j (\bar{y}_j - m_j). \quad (4)$$

To model m_j , we introduce the link function $g(m_j) = \mathbf{x}_j^T \boldsymbol{\beta}$, where $\mathbf{x}_j^T = (x_{j1}, \dots, x_{jp})$ are the vectors containing the auxiliary information at the area level, $\boldsymbol{\beta}^T = (\beta_1, \dots, \beta_p)$ is the regression parameter, and $g(\cdot)$ is a monotone increasing function. In case of canonical link, i.e. when $g \equiv (b')^{-1}$, $g'(m_j) = V^{-1}(m_j)$. It should be noted that unlike the HB model, no priors are assumed for the $p + 1$ hyperparameters, λ and $\boldsymbol{\beta}$, but they have to be estimated from the marginal distribution of $\bar{y}_j$ as is usual in an EB approach.

According to Sarkar and Ghosh (1998), the vector of parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}, \lambda)^T$ is estimated as solution of the generalized estimating equations (GEE) given by

$$\mathbf{S}_k(\boldsymbol{\psi}) = \frac{1}{k} \sum_{j=1}^k \mathbf{s}_j(\boldsymbol{\psi}) = \mathbf{0}. \quad (5)$$

²For example, in the binomial case we have $v_0 = 0, v_1 = 1, v_2 = -1$ and $\phi_j = 1, \forall j$.

In the above

$$\mathbf{s}_j(\boldsymbol{\psi}) = \begin{pmatrix} \omega_j \frac{(\bar{y}_j - m_j) \mathbf{x}_j}{V[m_j(\beta)]g'[m_j(\beta)]} \\ n_j(\bar{y}_j - \bar{y})^2 - n_j(1 - n_j/n)\omega_j^{-1}V(m_j) - n_j(m_j - \bar{m})^2 \end{pmatrix} \quad (6)$$

denoting $\bar{y} = \sum_{j=1}^k \bar{y}_j n_j / n$, $\bar{m} = \sum_{j=1}^k m_j n_j / n$ and $\omega_j = \omega_j(\lambda) = \frac{(\lambda - v_2)n_j \phi_j}{\lambda + \phi_j n_j}$. The consistency of $\hat{\boldsymbol{\psi}}$, solution³ of (5), has been proved by Sarkar and Ghosh (1998), while we will show their *asymptotic normality*. All the asymptotics results of this paper hold for $k \rightarrow \infty$ and bounded n_j , since in small area estimation problems we usually deal with a large number of domains, where most of them have small sample sizes.

Recalling (4), the EB estimator of γ_j , for $j = 1, \dots, k$, is now

$$e_j^{EB} = \bar{y}_j - f_j \hat{B}_j(\bar{y}_j - \hat{m}_j) \quad (7)$$

where $\hat{m}_j$ and $\hat{B}_j$ are obtained using the GEE estimates.

3. The unconditional mean squared error of prediction

The UMSEP for the EB estimator of the j -th population mean, is

$$\text{UMSEP}_j = E[e_j^{EB} - \gamma_j]^2 = E[(e_j^{EB} - e_j^B) + (e_j^B - \gamma_j)]^2. \quad (8)$$

Since $E(\gamma_j | \bar{y}_j) = e_j^B$,

$$E[(e_j^{EB} - e_j^B)(e_j^B - \gamma_j)] = E\{(e_j^{EB} - e_j^B)[e_j^B - E(\gamma_j | \bar{y}_j)]\} = 0. \quad (9)$$

Hence, the UMSEP is given by

$$\text{UMSEP}_j = E(e_j^B - \gamma_j)^2 + E(e_j^{EB} - e_j^B)^2, \quad (10)$$

where the first term is the mean squared error of the Bayes estimator if the vector of parameters $\boldsymbol{\psi}$ were known, while the second is the correction term that accounts for the estimation of $\boldsymbol{\psi}$. We will show that the second component tends to 0 as $k \rightarrow \infty$, and an intuitive explanation of this fact is the consistency of $\hat{\boldsymbol{\psi}}$ as an estimator of $\boldsymbol{\psi}$.

First we compute

$$E(e_j^B - \gamma_j)^2 = f_j^2 \frac{V(m_j)}{\lambda - v_2} \{B_j^2 + \frac{\lambda N_j}{\phi_j} [\frac{(1 - B_j)^2}{1 - f_j} + \frac{1}{f_j}]\}. \quad (11)$$

Finding this expression requires heavy algebra, and the details can be found from the authors. The correction term takes the form

$$E(e_j^{EB} - e_j^B)^2 = f_j^2 E[\hat{B}_j(\bar{y}_j - \hat{m}_j) - B_j(\bar{y}_j - m_j)]^2. \quad (12)$$

Setting $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}, \hat{\lambda})^T$, $t_j(\boldsymbol{\psi}) = B_j(\bar{y}_j - m_j)$, $t_j(\hat{\boldsymbol{\psi}}) = \hat{B}_j(\bar{y}_j - \hat{m}_j)$, we have

$$\begin{aligned} E(e_j^{EB} - e_j^B)^2 &= f_j^2 E[t_j(\hat{\boldsymbol{\psi}}) - t_j(\boldsymbol{\psi})]^2 = f_j^2 E[(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})^T \times \frac{\partial}{\partial \boldsymbol{\psi}} t_j(\boldsymbol{\psi}) + o(\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\|)]^2 \\ &\doteq f_j^2 E[\text{tr}\{\mathbf{A}_j(\boldsymbol{\psi})[(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})^T]\}] \doteq f_j^2 \text{tr}\{E[\mathbf{A}_j(\boldsymbol{\psi})]E[(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi})^T]\}, \end{aligned}$$

³The equations (5) can only be numerically solved.

where

$$E[\mathbf{A}_j(\boldsymbol{\psi})] = E\left[\frac{\partial}{\partial \boldsymbol{\psi}} t_j(\boldsymbol{\psi})\right]\left[\frac{\partial}{\partial \boldsymbol{\psi}} t_j(\boldsymbol{\psi})\right]^T = \begin{pmatrix} B_j^2 \frac{\mathbf{x}_j \mathbf{x}_j^T}{[g'(m_j)]^2} & \mathbf{0}_{p \times 1} \\ \mathbf{0}_{1 \times p} & \frac{n_j \phi_j V(m_j)}{(\lambda + n_j \phi_j)^3 (\lambda - v_2)} \end{pmatrix}. \quad (13)$$

By Delta method⁴ it can be shown, under mild assumptions, the asymptotic normality of $\hat{\boldsymbol{\psi}}$ so that we can approximate the variance-covariance matrix of $\hat{\boldsymbol{\psi}}$ by the inverse of its asymptotic Fisher information matrix, as $k \rightarrow \infty$. Hence, having in mind (13), the correction term (12) can be approximated by

$$E(e_j^{EB} - e_j^B)^2 \doteq f_j^2 \text{tr}\{E[\mathbf{A}_j(\boldsymbol{\psi})]k^{-1}[\mathbf{U}(\boldsymbol{\psi})\mathbf{W}^{-1}(\boldsymbol{\psi})\mathbf{U}^T(\boldsymbol{\psi})]^{-1}\}, \quad (14)$$

where

$$\mathbf{W}(\boldsymbol{\psi}) = \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{j=1}^k \begin{pmatrix} \omega_j V^{-1}(m_j)[g'(m_j)]^{-2} \mathbf{x}_j \mathbf{x}_j^T & \mathbf{0}_{p \times 1} \\ \mathbf{0}_{1 \times p} & n_j^2 \text{Var}(\bar{y}_j - \bar{y})^2 \end{pmatrix} \quad (15)$$

and

$$\mathbf{U}(\boldsymbol{\psi}) = \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{j=1}^k \begin{pmatrix} \omega_j V^{-1}(m_j)[g'(m_j)]^{-2} \mathbf{x}_j \mathbf{x}_j^T & \mathbf{0}_{p \times 1} \\ \xi_j [n_j^* \frac{n_j}{\omega_j} b'''(m_j) + 2n_j m_j^*] & -n_j^* \frac{V(m_j)}{(\lambda - v_2)^2} \frac{(n_j \phi_j + v_2)}{\phi_j} \end{pmatrix} \quad (16)$$

with $n_j^* = (1 - n_j/n)$, $m_j^* = m_j - \bar{m}$ and $\xi_j = \frac{\mathbf{x}_j^T}{g'(m_j)}$, for $j = 1, \dots, k$.

4. Results

By (10), (11) and (14), we have obtained a second-order approximation of the UMSEP for the EB predictors of small area means based on NEF-QVF superpopulation models. It is important to observe that, to guarantee the validity of these results, the number of small areas should be large while it does not matter if the sample size of some areas is zero. We developed a bias correction technique to estimate the UMSEP, which is not reported here.

References

- Ghosh M. and Meeden G. (1986) Empirical Bayes estimation in finite population sampling, *Journal of the American Statistical Association*, 81, 1058–1062.
- Ghosh M. and Rao J.N.K. (1994) Small area estimation: An appraisal, *Statistical Science*, 9, 55–93.
- McCullagh P. and Nelder J.A. (1989) *Generalized Linear Models (2nd edition)*, Chapman and Hall, London.
- Morris C. (1982) Natural exponential families with quadratic variance functions, *The Annals of Statistics*, 10, 65–80.
- Sarkar S. and Ghosh M. (1998) Empirical Bayes estimation of local area means for NEF-QVF superpopulations, *Sankhyā, B*, 60, 464–487.

⁴The proof is available from the authors.