SMALL AREA ESTIMATION: AN APPLICATION OF A FLEXIBLE FAY-HERRIOT METHOD

A. K. Wanjoya[1], N. Torelli[2] and G. Datta[3]
[1,2]Department of Statistics, University of Padua, Padua, Italy
[3]Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Kenya
[1]Statistics Department, University of Georgia, Athens, USA
E-mail: awanjoya@fsc.jkuat.ac.ke

Abstract
The importance of small area estimation in survey sampling is increasing, due to the growing demand for reliable small area estimation from both public and private sectors. In this paper, we address the important issue of using statistical modeling techniques to compute more reliable small area estimates. The main aim is to assess the use of a flexible methodology for small area estimation. We formulate a new flexible small area model by incorporating a tuning (index) parameter into the standard area-level (Fay-Herriot) model. We achieve this using a combination of two methods namely, empirical Bayes (EB) approach and hierarchical Bayes (HB) approach. Our results suggest that the proposed model can be seen as advancement over the standard Fay-Herriot model. The novelty here is that we have developed a flexible way to handle random effects in small area estimation. The Implementation of the proposed model is only mildly more difficult than the Fay-Herriot model. We have obtained results for both EB approach and the HB approach. Compared with the corresponding HB procedure, the EB approach saves a tremendous computing time and is very simple to implement.

Key words: Area-level, empirical Bayes, Fay-Herriot model, hierarchical Bayes, small area

1.0 Introduction

In recent years, the statistical technique of small area estimation (SAE) has been a very hot topic, and there is an ever-growing demand for reliable estimates of small area populations of all types. Reliable estimates of the population of small areas are important for several reasons. These estimates are used for, among other things, determination of state funding allocations, and determination of exact boundaries for schools and voting districts, administrative planning, disease mapping, marketing guidance and as data for detailed descriptive and analytical studies for cities (Bryan, 1999).

According to Pfeffermann (2002), the problem of small area estimation is twofold. First is the fundamental question of how to produce reliable estimates of characteristics of interest, (means, counts, quantiles, etc.) for small areas or domains, based on very small samples taken from these areas. The second related question is how to assess the estimation error. Note in this respect that except in rare cases, sampling designs and in particular sample sizes are chosen in practice so as to produce reliable estimates for aggregates of small areas such as geographic regions or demographic groups. Budget and other constraints usually prevent the allocation of sufficiently large samples to each of the small areas. Also, it is often the case that domains of interest are only specified after the survey has already been designed and carried out. Having only a small sample (and possibly an empty sample) in a given area, the only possible solution to the estimation problem is to *borrow information* from other related data sets. Potential data sources can be divided into two broad categories: data measured for the characteristics of interest in other 'similar' areas or data measured for the characteristics of interest on previous occasions.

The methods used for SAE can be divided accordingly by the related data sources they employ or by type of inference: 'design based', 'model dependent' (with subdivision into the frequentist and Bayesian approaches), or the combination of the two. Given the growing use of small area statistics and their immense importance, it is imperative to develop efficients tools or models for small area estimation and ascertainment of their goodness of fit taking into account relationships between small areas.

In this paper, we address the important issue of using statistical modelling techniques to compute more reliable small area estimates. The main aim is to assess the use of a flexible methodology for small area estimation. We formulate a new flexible smallarea model by incorporating a tuning (index) parameter into the standard area-level (Fay-Herriot) model. We achieve this using a combination of two methods namely, empirical Bayes (EB) approach and hierarchical Bayes (HB) approach. To that end, after describing the small area model-based methods in section 2, we outline the proposed small area flexible model in section 3. In section

4 we report results of estimation of median incomes of four person families using US survey data. Finally section 5 gives some concluding remarks.

2.0 Current Model-Based Approaches to Small Area Estimation

Small area estimation is one of the few fields in survey sampling where it is widely recognized that the use of model dependent inference is often inevitable. The model-based approach to small area estimation permits validation of models from sample data.Ghosh and Rao (1994), Rao (2003) and Torabi and Rao (2008) classify small area models into two types:

$$\theta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \nu_i \qquad \text{.......(1)}$$

In this Fay-Herriot model (1) (Fay & Herriot, 1979), area-specific auxiliary data $\boldsymbol{x}_i$ (administrative records,census data) are available for the areas $i = 1,2, \ldots, m$. The population small area total Y_i, or some function $a_i = g(Y_i)$, is assumed to be related to $\boldsymbol{x}_i$ through the linear model (1). The ν_i's are assumed to be normally distributed, random, uncorrelated small area effects, with mean zero and variance σ_ν^2. $\boldsymbol{\beta}$ represents the vector of regression parameters. The second type of model is as follows:

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \nu_i + \epsilon_{ij} \qquad \text{.......(2)}$$

This model is appropriate for continuous variables y. In model (2), unit-specific auxiliary data $\boldsymbol{x}_{ij}$ are again available for the areas $i = 1,2, \ldots, m$, where $j = 1,2, \ldots, N_i$ and N_i represents the number of population units in the i-th area. The unit y-values, y_{ij}, are assumed to be related to the auxiliary values $\boldsymbol{x}_{ij}$ through the nested error regression model (2) where $\nu_i \sim \mathcal{N}(0, \sigma_\nu^2)$ and $\epsilon_{ij} \sim thcalN(0, \sigma_\epsilon^2)$ ($\sim$ denotes independent and identically distributed as), ν_i and ϵ_{ij} are assumed to be mutually independent. $\boldsymbol{\beta}$ again represents the vector of regression parameters.

Rao (2003) and Torabi and Rao (2008) further asserts that in the case of models (1), direct survey estimators $\hat{Y}_i$ are available whenever the sample sizes $n_i \geq 1$ and it can be assumed that

$$\hat{\theta}_i = \theta_i + \epsilon_i \qquad \text{.......(3)}$$

where $\hat{\theta}_i = g(\hat{Y}_i)$ and the sampling errors $\epsilon_i \sim \mathcal{N}(0, \psi_i)$. Then, when model (3) is combined with model (1), we have

$$\hat{\theta}_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \nu_i + \epsilon_i \qquad \text{.......(4)}$$

which is a special case of the general linear mixed model. Note that model (4) involves design variables, ϵ_i, as well as model-based random variables ν_i.

According to Rao (1999), "The success of small area estimation largely depends on getting good auxiliary information ($\boldsymbol{x}_i$) that leads to small area model variance σ_v^2 relative to ψ_i."

A variety of approaches such as (empirical-) best linear unbiased prediction (E-BLUP), empirical Bayes (EB) and hierarchical Bayes (HB) are commonly used in model-based small area estimation. The techniques of maximum likelihood (ML), restricted maximum likelihood (REML), penalized quasi-likelihood, etc. have been utilized for estimates of the model-based estimators. Details of theoretical techniques for the estimation of the parameters for different types of small area models are discussed by Rao (2003; and references therein).

3.0 Description of the Proposed Model

3.1 Proposed Model

In the proposed model we assume that there exists a direct survey estimator y_i for the small area parameter θ_i such that

$$y_i = \theta_i + e_i,$$

and

$$\theta_i = x_i^T \beta + \delta_i \nu_i, \quad i = 1, \dots, m$$

where m is the number of small areas, $\beta = (\beta_1, \dots, \beta_p)'$ is $p \times 1$ vector of regression coefficients, and the ν_i's are area-specific random effects assumed to be independent and identically distributed (iid) with $E(\nu_i) = 0$ and $var(\nu_i) = A$. $\delta_1, \dots, \delta_m$ are iid Bernoulli random variables. $\nu_1, \dots, \nu_m$ and $\delta_1, \dots, \delta_m$ are assumed to be independent.

Given that $\delta_i = 1$, $\nu_i \sim \mathcal{N}(0, A)$, $pr(\delta_i = 1) = p$ and assuming that A, p, and β are known, the Bayes predictor of θ_i becomes:

$$\begin{aligned} \hat{\theta}_i^B &= E(\theta_i|\boldsymbol{y}) = x_i^T \beta + E[\delta_i \nu_i|\boldsymbol{y}] = x_i^T \beta + E[E(\delta_i \nu_i|\delta_i, \boldsymbol{y})|\boldsymbol{y}] \\ &= x_i^T \beta + E[\nu_i|\delta_i = 1, \boldsymbol{y}] \cdot P[\delta_i = 1|\boldsymbol{y}] \end{aligned}$$

On observing that

$$y_i|\nu_i, \delta_i = 1, \beta \sim \mathcal{N}(x_i^T \beta + \nu_i, D_i); \quad \nu_i|\delta_i = 1 \sim \mathcal{N}(0, A);$$

and

$$\nu_i|\delta_i = 1, \beta, \boldsymbol{y} \sim \mathcal{N}\left(\frac{A}{A + D_i}(y_i - x_i^T \beta), \frac{AD_i}{A + D_i}\right);$$

We have

$$\hat{\theta}_i^B = x_i^T\beta + \frac{A}{A+D_i}(y_i - x_i^T\beta)\cdot P(\delta_i = 1|\boldsymbol{y}, p, \beta, A)$$
$$= x_i^T\beta + \frac{A}{A+D_i}(y_i - x_i^T\beta)\cdot \hat{p}_i(p, \beta, A);$$

the probability $\hat{p}_i(p, \beta, A)$ is derived by observing that

$$P(\delta_i = 1|\boldsymbol{y}, \delta, A) = P(\delta_i = 1|y_i, \delta, A) = \frac{P(\delta_i = 1, y_i)}{f(y_i)}$$

$$= \frac{f(y_i|\delta_i = 1)P(\delta_i = 1)}{f(y_i|\delta_i = 1)P(\delta_i = 1) + f(y_i|\delta_i = 0)P(\delta_i = 0)}.$$

But

$$y_i|\delta_i = 1 \sim \mathcal{N}(x_i^T\beta, A + D_i) \quad \text{and} \quad y_i|\delta_i = 0 \sim \mathcal{N}(x_i^T\beta, D_i).$$

Therefore,

$\hat{p}_i(p, \beta, A)$

$$= \frac{\frac{1}{\sqrt{2\pi(A+D_i)}} exp\left(-\frac{(y_i - x_i^T\beta)^2}{2(A+D_i)}\right) \times p}{\frac{1}{\sqrt{2\pi(A+D_i)}} exp\left(-\frac{(y_i - x_i^T\beta)^2}{2(A+D_i)}\right) \times p + \frac{1}{t2\pi D_i} exp\left(-\frac{(y_i - x_i^T\beta)^2}{2D_i}\right) \times (1-p)}$$

Hence the marginal density of Y_i, $f(y_i)$, is:

$$f(y_i) = \frac{p}{\sqrt{2\pi(A+D_i)}} \exp\left(-\frac{(y_i - x_i^T\beta)^2}{2(A+D_i)}\right) + \frac{(1-p)}{\sqrt{2\pi D_i}} \exp\left(-\frac{(y_i - x_i^T\beta)^2}{2D_i}\right) \quad \text{................(5)}$$

The Empirical Bayes predictor $(\hat{\theta}_i^{EB}(\hat{\beta}, \hat{p}, \hat{A}; y_i))$ of θ_i can be obtained by estimating the parameters β, p and A from the marginal distribution of $Y_1, \dots, Y_m$:

$$f(\boldsymbol{y}|\beta, p, A) = \prod_{i=1}^{m}\left[\frac{p}{\sqrt{2\pi(A+D_i)}}\exp\left(-\frac{(y_i - x_i^T\beta)^2}{2(A+D_i)}\right) + \frac{(1-p)}{\sqrt{2\pi D_i}}\exp\left(-\frac{(y_i - x_i^T\beta)^2}{2D_i}\right)\right] \quad \ldots\ldots\ldots\ldots\ldots(6)$$

Note that $A = 0$ will lead to $\hat{p} = 0$. On the other hand, $p = 0$ will make the estimation of A impossible. So we shall assume that $p > 0$ and $A > 0$.

A hierarchical Bayesian approach is developed to estimate parameters of the proposed model, with the implementation carried out by Markov Chain Monte Carlo (MCMC) techniques. This requires generation of samples from the full conditional distributions given in the appendix.

3.2 Empirical Comparisons

We use the following four criteria to compare the estimates obtained via the standard Fay-Herriot Model and the proposed model. Suppose e_{iTR} denotes the true value for the ith small area, and e_i is any estimate of e_{iTR}, $i = 1, \cdots, m$. Then

Average relative bias (ARB) = $\frac{1}{m}\sum_{i=1}^{m}\left|\frac{e_i - e_{iTR}}{e_{iTR}}\right|$

Average squared relative bias (ASRB) = $\frac{1}{m}\sum_{i=1}^{m}\left(\frac{e_i - e_{iTR}}{e_{iTR}}\right)^2$

Average absolute bias (AAB) = $\frac{1}{m}\sum_{i=1}^{m}|e_i - e_{iTR}|$

and

Average squared deviation (ASD) = $\frac{1}{m}\sum_{i=1}^{m}(e_i - e_{iTR})^2$

4.0 Data Analysis

In this section, we report findings after using the proposed model to analyse the Median Income survey data set for the 50 states in United States (US) and District of Columbia (DC). This survey data set was collected by Bureau of Economic Analysis of the U.S. Department of Commerce. Our findings after comparing the estimates according to four criteria introduced in section 3 are summarized in four tables. We implement the model via the empirical Bayes (EB) approach as well as the hierarchical Bayes (HB) approach.

4.1 Empirical Bayes Approach

Table 1: Empirical Comparison of EB Estimates under Fay-Herriot (FH) and Proposed Model (PM)

Model	Average relative deviation	Average squared relative deviation	Average absolute deviation	Average squared deviation
FH	843059.09	0.00206	724.81	0.0358
PM	688768.47	0.00178	675.46	0.0339

Table 2: Empirical Comparison of EB Estimates under Fay-Herriot (FH) and Proposed Model (PM)

Model	Average relative deviation	Average squared relative deviation	Average absolute deviation	Average squared deviation
p = 0.1				
FH	140151.26	0.000339	219.20	0.0109
PM	113878.44	0.000257	171.85	0.0081
p = 0.25				
FH	380009.2	0.00089	486.82	0.0235
PM	372436.78	0.00082	452.45	0.0217
p = 0.50				
FH	859917.39	0.00217	757.50	0.0374
PM	732942.42	0.00188	702.07	0.0348
p = 0.75				
FH	925853.67	0.00216	746.49	0.0364
PM	893110.28	0.00208	729.75	0.0359

Tables 1 and 2 report the figures for different estimates. It is clear from both tables that the estimates obtained by the proposed model improve substantially over the standard Fay-Herriot model estimates. The corresponding percentage improvements range from 5% to 26%.

4.2 Hierarchical Bayes Approach

The Models were fitted in *R*, using two parallel Markov Chain Monte Carlo chains of 8,000 iterations following burn-in of 2,000. Very intensive computation was involved for the HB model (e.g. 27 hours on a 2.4 GHz processor with 2Gb RAM). Satisfactory convergence was confirmed using the Gelman and Rubin convergence statistic. Samples of 4,000 from the posterior distributions were obtained from a 1:4 thinning of the combined chains and summarised to provide estimates.

Table 3: Empirical Comparison of HB Estimates under Fay-Herriot (FH) and Proposed Model (PM)

Model	Average relative deviation	Average squared relative deviation	Average absolute deviation	Average squared deviation
FH	806868.51	0.00198	711.813	0.0352
PM	658044.94	0.00166	670.047	0.0334

Table 4: Empirical Comparison of HB Estimates under Fay-Herriot (FH) and Proposed Model (PM)

Model	Average relative deviation	Average squared relative deviation	Average absolute deviation	Average squared deviation
p = 0.1				
FH	173940.86	0.00043	280.08	0.0139
PM	146989.92	0.00036	229.85	0.0114
p = 0.25				
FH	392605.54	0.00093	497.554	0.0241
PM	385753.19	0.00088	460.073	0.0221
p = 0.50				
FH	869608.51	0.00219	761.92	0.0376
PM	740403.12	0.00192	704.99	0.0351
p = 0.75				
FH	899831.83	0.00209	739.34	0.0361
PM	861020.57	0.00200	727.47	0.0355

The statistical results of the hierarchical Bayesian approach are presented in table 3 and 4. From these tables, we see that we obtain reasonably better estimates using the proposed model. This finding is supported by all the four comparison criteria employed in the analysis.
Our results of fitting the proposed flexible model and the standard Fay-Herriot model to the Median Income survey data set suggest that the proposed model can be seen as advancement over the standard Fay-Herriot model. The proposed model enables us to obtain a higher quality of the estimates.

5.0 Conclusion

To conclude, based on the results of fitting the proposed flexible model and the standard Fay-Herriot model to the Median Income survey data set for the 50 states in United States (U.S.) and District of Columbia (DC), the proposed model appears to be a good alternative to the standard Fay-Herriot model and we can tentatively recommend the use of the proposed model. The novelty here is that we have developed a flexible way to handle random effects in small area estimation. The implementation of the proposed model is only mildly more difficult than the Fay-Herriot model. We have obtained results for both the empirical Bayes (EB) approach and the hierarchical Bayes (HB) approach. Compared with the corresponding HB procedure, the EB approach saves a tremendous computing time and is very simple to implement. An advantage of the HB approach is that the inferences about the parameters are "exact" unlike the EB approach. The HB approach will automatically take into account the uncertainties associated with unknown parameters. However, it does require the specification of prior distributions. It may be a rewarding topic for future research to investigate whether this approach can be applied to situations where the response variable is not continuous and normally distributed.

Appendix: Full Conditionals

Bayesian Formulation:

$y_i|\theta_i \sim \mathcal{N}(\theta_i, D_i), \quad \theta_i \sim \mathcal{N}(\boldsymbol{x}_i^T\beta, A), \quad (i = 1,2,\cdots,m).$

We apply Gibbs sampling method to generate samples from the full conditional distributions of the proposed model:

Conditional on the parameters β and A,

$$\theta_i|\boldsymbol{y},\beta,A \sim \mathcal{N}\left(\frac{D_i x_i^T\beta + Ay_i}{A + D_i}, \frac{AD_i}{A + D_i}\right)$$

Conditional on the parameters θ and A,

$$\beta|\boldsymbol{y},\boldsymbol{\theta},A \sim \mathcal{N}((X^TX)^{-1}X^T\theta, A(X^TX)^{-1})$$

Conditional on the parameters θ and β,

$$\pi(A|\beta,\boldsymbol{\theta},\boldsymbol{y}) \quad \propto \quad \exp\left(-\sum_{i=1}^{m}\frac{(\theta_i - X_i^T\beta)^2}{2A}\right)A^{-\frac{m}{2}}\frac{1}{(A + \bar{D})^2}$$

Conditional distribution of β given θ, A and $\boldsymbol{y}$ is

$$\beta|\boldsymbol{y},\delta,p,A \quad \sim \quad \mathcal{MVN}(H^{-1}g, H^{-1})$$

where $H = \sum_{i=1}^{m}\left\{\frac{\delta_i}{A+D_i} + \frac{(1-\delta_i)}{D_i}\right\}x_i x_i^T$, $g = \sum_{i=1}^{m}\left\{\frac{\delta_i}{A+D_i} + \frac{(1-\delta_i)}{D_i}\right\}x_i y_i$, $x_i = (1, x_{i1})^T$

,

Conditional on the parameters θ, β, p and $\boldsymbol{y}$,

$$814\; pr(\delta_i = 1|p, \beta, A, \boldsymbol{y}) = \frac{p}{p + (1-p)\sqrt{\frac{A + D_i}{D_i}}\, exp\left\{-\frac{(y_i - x_i^T\beta)^2}{2}\left(\frac{1}{D_i} - \frac{1}{A + D_i}\right)\right\}}$$

Acknowledgements

The first author gratefully acknowledges the PhD grant from Fondazione Cassa di Risparmio di Padova è Rovigo (CARIPARO), Italy.

References

Bryan T. (1999). Small area population estimation technique using administrative records, and evaluation of results with loss functions and optimization criteria. Available from: http:/www.census.gov/www/cob/bg.html

Fay R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **85**, pp 398-409.

Ghosh M. and Rao J. N. K. (1994). Small Area Estimation: An appraisal. *Statistical Science,* **9**(1), pp. 55-76.

Pfeffermann D. (2002). Small area estimation- new developments and directions. *International Statistical Review,* **70**, pp 125-143.

Rao J. N. K. (1999). Current trends in sample survey theory and methods. *Indian Journal of Statistics,* **61**, pp 16-22.

Rao J. N. K. (2003). *Small Area Estimation*. John Wiley and Sons, Inc., Hoboken, New Jersey.

Torabi M. and Rao J. N. K. (2008). Small area estimation under a two-level model. *Survey Methodology*, **34**, pp 11-17.