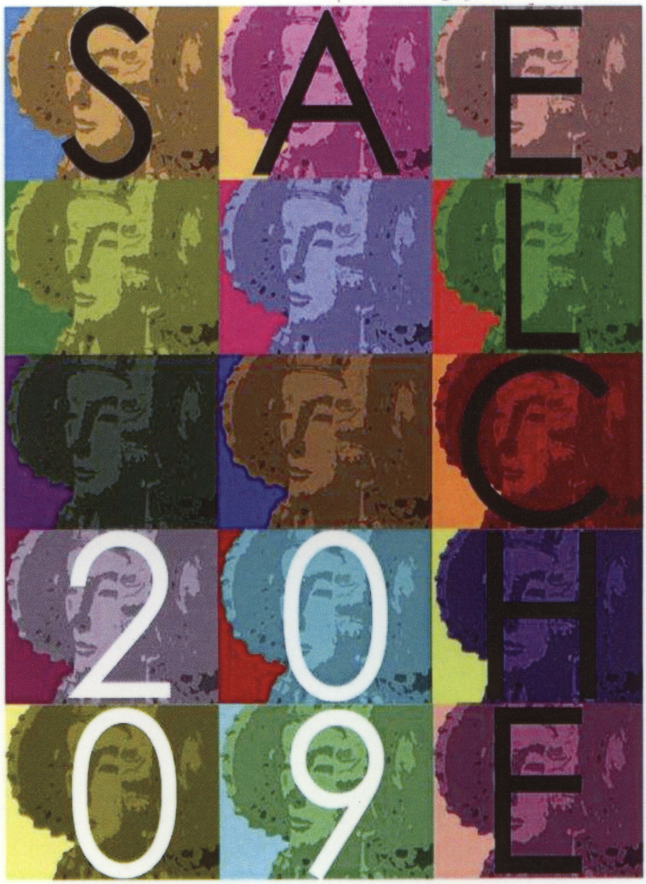
Kusman Sadik
Dept. Statistika IPB
HLm : 55



Plenary Speakers

Ana Fernández Militino
Universidad Pública de Navarra
Spain

Malay Ghosh
University of Florida
U.S.A.

Francisco Hernández Jiménez
Instituto Nacional de Estadística
Spain

Partha Lahiri
University of Maryland
U.S.A.

Sharon Lohr
Arizona State University
U.S.A.

Pedro Silva
Brazilian Central Statistical Office
Brazil

Li-Chun Zhang
Statistics Norway
Norway

SAE2009Conference on Small Area Estimation

Organized by

Universidad Miguel Hernández de Elche
Instituto Nacional de Estadística
Ayuntamiento de Elche

Elche, Alicante, Spain
June 29 - July 1, 2009

http://cio.umh.es/sae2009

M2A. Small area estimation for business surveys

29/06,16:00
ChairPerson: Susana Rubin-Bleuer
Room: A

M2B. Bayesian cross-sectional modelling

29/06,16:00
ChairPerson: Enrico Fabrizi
Room: B

M2C. Census, demography and benchmarking

29/06,16:00
ChairPerson: Ralf Münnich
Room: C

SAE2009 Conference on Small Area Estimation
June 29 - July 01, 2009, Elche, Spain

Hierarchical Bayes Estimation Using Time Series and Cross-sectional Data: A Case of Per-capita Expenditure in Indonesia

Kusman Sadik and Khairil Anwar Notodiputro

Department of Statistics, Bogor Agricultural University / IPB
Jl. Raya Dramaga, Bogor, Indonesia 16680

Abstract

In Indonesia, there is a growing demand for reliable small area statistics in order to assess or to put into policies and programs. Sample survey data provide effective reliable estimators of totals and means for large area and domains. But it is recognized that the usual direct survey estimator performing statistics for a small area, have unacceptably large standard errors, due to the circumstance of small sample size in the area. The primary source of data for this paper is the National Socio-economic Survey (Susenas), a survey which is conducted every year in Indonesia. However, the estimation of Susenas village per-capita expenditure is unreliable, due to the limited number of observations per village. Hence, it is important to improve the estimates. We proposed a hierarchical Bayes (HB) method using a time series generalization of a widely used cross-sectional model in small area estimation. Generalized variance function (GVF) is used to obtain the estimates of sampling variance.

Key words: Linear mixed model, Hierarchical Bayes, posterior predictive assessment, generalized variance function, block diagonal covariance, Kalman filter, state space model.

1. Introduction

The problem of small area estimation is how to produce reliable estimates of area (domain) characteristics, when the sample sizes within the areas are too small to warrant the use of traditional direct survey estimates. Sample survey data provide effective reliable estimators of totals and means for large areas and domains. But it is recognized that the usual direct survey estimators performing statistics for a small area, have unacceptably large standard errors, due to the circumstance of small sample size in the area. In fact, sample sizes in small areas are reduced, due to the circumstance that the overall sample size in a survey is usually determined to provide specific accuracy at a macro area level of aggregation, that is national territories, regions ad so on (Datta and Lahiri, 2000).

Demand for reliable small area statistics has steadily increased in recent years which prompted considerable research on efficient small area estimation. Direct small area estimators from survey data fail to borrow strength from related small areas since they are based solely on the sample data associated with the corresponding areas. As a result, they are likely to yield unacceptably large standard errors unless the sample size for the small area is reasonably large(Rao, 2003). Small area efficient statistics provide, in addition of this, excellent statistics for local estimation of population, farms, and other characteristics of interest in post-censual years.

2. Indirect Estimation in Small Area

A domain (area) is regarded as large (or major) if domain-specific sample is large enough to yield direct estimates of adequate precision. A domain is regarded as small if the domain-specific sample is not large enough to support direct estimates of adequate precision. Some other terms used to denote a domain with small sample size include local area, sub-domain, small subgroup, sub-province, and minor domain. In some applications, many domains of interest (such as counties) may have zero sample size.

In making estimates for small area with adequate level of precision, it is often necessary to use indirect estimators that borrow strength by using thus values of the variable of interest, y, from related areas and/or time periods and thus increase the effective sample size. These values are

brought into the estimation process through a model (either implicit or explicit) that provides a link to related areas and/or time periods through the use of supplementary information related to *y*, such as recent census counts and current administrative records (Pfeffermann 2002; Rao 2003).

Methods of indirect estimation are based on explicit small area models that make specific allowance for between area variation. In particular, we introduce mixed models involving random area specific effects that account for between area variation beyond that explained by auxiliary variables included in the model. We assume that $\theta_i = g(\bar{Y}_i)$ for some specified $g(.)$ is related to area specific auxiliary data $\mathbf{z}_i = (z_{1i}, \ldots, z_{pi})^T$ through a linear model

$$\theta_i = \mathbf{z}_i^T\boldsymbol{\beta} + b_i v_i, \qquad i = 1, \ldots, m$$

where the b_i are known positive constants and $\boldsymbol{\beta}$ is the px1 vector of regression coefficients. Further, the v_i are area specific random effects assumed to be independent and identically distributed (iid) with

$$E_m(v_i) = 0 \text{ and } V_m(v_i) = \sigma_v^2\ (\geq 0), \text{ or } v_i \sim \text{iid}\ (0, \sigma_v^2)$$

3. General Linear Mixed Model

Datta and Lahiri (2000), and Rao(2003) considered a general linear mixed model (GLMM) which covers the univariate unit level model as special cases:

$$\mathbf{y}^P = \mathbf{X}^P\boldsymbol{\beta} + \mathbf{Z}^P\mathbf{v} + \mathbf{e}^P$$

Hence $\mathbf{v}$ and $\mathbf{e}^P$ are independent with $\mathbf{e}^P \sim \mathbf{N}(\mathbf{0}, \sigma^2\boldsymbol{\psi}^P)$ and $\mathbf{v} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{D}(\boldsymbol{\lambda}))$, where $\boldsymbol{\psi}^P$ is a known positive definite matrix and $\mathbf{D}(\boldsymbol{\lambda})$ is a positive definite matrix which is structurally known except for some parameters $\boldsymbol{\lambda}$ typically involving ratios of variance components of the form σ_i^2/σ^2. Further, $\mathbf{X}^P$ and $\mathbf{Z}^P$ are known design matrices and $\mathbf{y}^P$ is the N x 1 vector of population y-values. The GLMM form :

$$\mathbf{y}^P = \begin{bmatrix}\mathbf{y}\\ \mathbf{y}^*\end{bmatrix} = \begin{bmatrix}\mathbf{X}\\ \mathbf{X}^*\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\mathbf{Z}\\ \mathbf{Z}^*\end{bmatrix}\mathbf{v} + \begin{bmatrix}\mathbf{e}\\ \mathbf{e}^*\end{bmatrix}$$

where the asterisk (*) denotes non-sampled units. The vector of small area totals (Y_i) is of the form $\mathbf{Ay} + \mathbf{Cy}^*$ with $\mathbf{A} = \oplus_{i=1}^m \mathbf{1}_{n_i}^T$ and $\mathbf{C} = \oplus_{i=1}^m \mathbf{1}_{N_i - n_i}^T$ where $\oplus_{i=1}^m \mathbf{A}_u = \text{blockdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_m)$.

We are interested in estimating a linear combination, $\mu = \mathbf{l}^T\boldsymbol{\beta} + \mathbf{m}^T\mathbf{v}$, of the regression parameters $\boldsymbol{\beta}$ and the realization of $\mathbf{v}$, for specified vectors, $\mathbf{l}$ and $\mathbf{m}$, of constants. For known $\boldsymbol{\delta}$, the BLUP (*best linear unbiased prediction*) estimator of μ is given by (Rao, 2003)

$$\tilde{\mu}^H = t(\boldsymbol{\delta}, \mathbf{y}) = \mathbf{l}^T\tilde{\boldsymbol{\beta}} + \mathbf{m}^T\tilde{\mathbf{v}} = \mathbf{l}^T\tilde{\boldsymbol{\beta}} + \mathbf{m}^T\mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$$

Model of indirect estimation, $\hat{\theta}_i = \mathbf{z}_i^T\boldsymbol{\beta} + b_i v_i + e_i$, $i = 1, \ldots, m$, is a special case of GLMM with block diagonal covariance structure. Making the above substitutions in the general form for the BLUP estimator of μ_i, we get the BLUP estimator of θ_i as:

$$\tilde{\theta}_i^H = \mathbf{z}_i^T\tilde{\boldsymbol{\beta}} + \gamma_i(\hat{\theta}_i - \mathbf{z}_i^T\tilde{\boldsymbol{\beta}}), \text{ where } \gamma_i = \sigma_v^2 b_i^2/(\psi_i + \sigma_v^2 b_i^2), \text{ and}$$

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\sigma_v^2) = \left[\sum_{i=1}^m \frac{\mathbf{z}_i\mathbf{z}_i^T}{\psi_i + \sigma_v^2 b_i^2}\right]^{-1}\left[\sum_{i=1}^m \frac{\mathbf{z}_i\hat{\theta}_i}{\psi_i + \sigma_v^2 b_i^2}\right]$$

4. Hierarchical Bayes for State Space Models

Many sample surveys are repeated in time with partial replacement of the sample elements. For such repeated surveys considerable gain in efficiency can be achieved by borrowing strength across both small areas and time. Their model consist of a sampling error model

$$\hat{\theta}_{it} = \theta_{it} + e_{it}, \quad t = 1, \ldots, T;\ i = 1, \ldots, m$$

$$\theta_{it} = \mathbf{z}_{it}^T\boldsymbol{\beta}_{it}$$

where the coefficients $\boldsymbol{\beta}_{it} = (\beta_{it0}, \beta_{it1}, \ldots, \beta_{itp})^T$ are allowed to vary cross-sectionally and over time, and the sampling errors e_{it} for each area i are assumed to be serially uncorrelated with mean 0 and variance ψ_{it}. The variation of $\boldsymbol{\beta}_{it}$ over time is specified by the following model:

$$\begin{bmatrix}\beta_{itj}\\ \beta_{ij}\end{bmatrix} = \mathbf{T}_j\begin{bmatrix}\beta_{i,t-1,j}\\ \beta_{ij}\end{bmatrix} + \begin{bmatrix}1\\ 0\end{bmatrix} v_{itj}, \quad j = 0,1,..., p$$

It is a special case of the general state-space model which may be expressed in the form

$$\mathbf{y}_t = \mathbf{Z}_t\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t; \qquad E(\boldsymbol{\varepsilon}_t) = 0, \quad E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^T) = \boldsymbol{\Sigma}t$$
$$\boldsymbol{\alpha}_t = \mathbf{H}_t\boldsymbol{\alpha}_{t-1} + \mathbf{A}\boldsymbol{\eta}_t; \qquad E(\boldsymbol{\eta}_t) = 0, \quad E(\boldsymbol{\eta}_t\boldsymbol{\eta}_t^T) = \boldsymbol{\Gamma}$$

where $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\eta}_t$ are uncorrelated contemporaneously and over time. The first equation is known as the *measurement equation*, and the second equation is known as the *transition equation*. This model is a special case of the general linear mixed model but the state-space form permits updating of the estimates over time, using the Kalman filter equations, and smoothing past estimates as new data becomes available, using an appropriate smoothing algoritm.

The vector $\boldsymbol{\alpha}_t$ is known as the *state vector*. Let $\tilde{\boldsymbol{\alpha}}_{\mathbf{t-1}}$ be the BLUP estimator of $\boldsymbol{\alpha}_{t-1}$ based on all observed up to time (t-1), so that $\tilde{\boldsymbol{\alpha}}_{\mathbf{t|t-1}} = \mathbf{H}\tilde{\boldsymbol{\alpha}}_{\mathbf{t-1}}$ is the BLUP of $\boldsymbol{\alpha}_t$ at time (t-1). Further, $\mathbf{P}_{t|t-1} = \mathbf{H}\mathbf{P}_{t-1}\mathbf{H}^T + \mathbf{A}\boldsymbol{\Gamma}\mathbf{A}^T$ is the covariance matrix of the prediction errors $\tilde{\boldsymbol{\alpha}}_{\mathbf{t|t-1}} - \boldsymbol{\alpha}_t$, where

$$\mathbf{P}_{t-1} = E(\tilde{\boldsymbol{\alpha}}_{\mathbf{t-1}} - \boldsymbol{\alpha}_{t-1})(\tilde{\boldsymbol{\alpha}}_{\mathbf{t-1}} - \boldsymbol{\alpha}_{t-1})^T$$

is the covariance matrix of the prediction errors at time (t-1). At time t, the predictor of $\boldsymbol{\alpha}_t$ and its covariance matrix are updated using the new data ($\mathbf{y}_t$, $\mathbf{Z}_t$). We have

$$\mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{\mathbf{t|t-1}} = \mathbf{Z}_t(\boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{\mathbf{t|t-1}}) + \boldsymbol{\varepsilon}_t$$

which has the linear mixed model form with $\mathbf{y} = \mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{\mathbf{t|t-1}}$, $\mathbf{Z} = \mathbf{Z}_t$, $\mathbf{v} = \boldsymbol{\alpha}_t - \tilde{\boldsymbol{\alpha}}_{\mathbf{t|t-1}}$, $\mathbf{G} = \mathbf{P}_{t|t-1}$ and $\mathbf{V} = \mathbf{F}_t$, where $\mathbf{F}_t = \mathbf{Z}_t\mathbf{P}_{t|t-1}\mathbf{Z}_t^T + \boldsymbol{\Sigma}_t$. Therefore, the BLUP estimator $\tilde{\mathbf{v}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\mathbf{y}$ reduces to

$$\tilde{\boldsymbol{\alpha}}_{\mathbf{t-1}} = \tilde{\boldsymbol{\alpha}}_{\mathbf{t|t-1}} + \mathbf{P}_{t|t-1}\mathbf{Z}_t^T\mathbf{F}_t^{-1}(\mathbf{y}_t - \mathbf{Z}_t\tilde{\boldsymbol{\alpha}}_{\mathbf{t|t-1}})$$

Based on the research conducted by Datta, Lahiri, and Maiti (2002), the model can be written as a hierarchical Bayes approach. Let $\hat{\theta}_{it}$ be the estimator of θ_{it} for the area *i*-th and *t*-th year (i = 1, 2, …, m; t = 1, 2, …, T). We shall consider the following hierarchical longitudinal model to improve on the estimate $\hat{\theta}_{iT}$ of θ_{iT}.

Level 1 : $\hat{\theta}_{it}|\theta_{it} \sim N(\theta_{it}, \psi_{it})$, where ψ_{it} is known
Level 2 : $\theta_{it}|\boldsymbol{\beta},\alpha_{it} \sim N(\mathbf{x}_{it}^T\boldsymbol{\beta} + \alpha_{it}, \tau^2)$
Level 3 : $\alpha_{it}|\alpha_{it-1} \sim N(\alpha_{it-1}, \sigma^2)$
Level 4 : $\boldsymbol{\beta}$ is flat, τ^2 and σ^2 is Gamma distribution

We assume the component $\alpha_{i0} = \alpha$ (i = 1, 2, …, m) is a fixed unknown constant, and ψ_{it} (i = 1, 2, …, m) are known fixed T x T positive definite matrices.

5. Case Study

Model of small area estimation can be applied to estimate average of households expenditure per month for the villages in Bogor county, East Java, Indonesia. Data that be used in this case study are data of Susenas (National Economic and Social Survey, BPS) 2001 to 2005.

Table 1. Design based and model based estimates of households expenditure per-month for the villages

Village	Design Based (Direct Estimator)		Model Based (Indirect Estimator)			
			EBLUP		Hierarchical Bayes	
	$\hat{\theta}_{iT}$	$mse(\hat{\theta}_{iT})$	$\hat{\theta}_{iT}^{P}$	$mse(\hat{\theta}_{iT}^{P})$	$\hat{\theta}_{iT}^{HB}$	$mse(\hat{\theta}_{iT}^{HB})$
1	5.71	0.132	5.74	0.127	6.12	0.120
2	5.07	0.119	4.75	0.118	5.74	0.081
3	4.65	0.090	4.96	0.083	5.28	0.067
4	5.98	0.142	5.55	0.124	6.15	0.131
5	5.50	0.139	5.16	0.116	5.46	0.121
6	5.52	0.161	4.84	0.145	4.16	0.132
7	4.89	0.102	4.61	0.097	4.15	0.086

Village	Design Based (Direct Estimator)		Model Based (Indirect Estimator)			
			EBLUP		Hierarchical Bayes	
	$\hat{\theta}_{iT}$	$mse(\hat{\theta}_{iT})$	$\hat{\theta}_{iT}^{P}$	$mse(\hat{\theta}_{iT}^{P})$	$\hat{\theta}_{iT}^{HB}$	$mse(\hat{\theta}_{iT}^{HB})$
8	5.06	0.093	5.25	0.067	4.50	0.047
9	6.02	0.114	5.75	0.061	6.47	0.046
10	6.29	0.106	6.47	0.123	5.69	0.065
11	8.49	0.186	9.07	0.167	9.01	0.178
12	6.61	0.140	5.69	0.091	7.02	0.076
13	9.45	0.204	9.51	0.205	9.01	0.196
14	8.40	0.162	8.33	0.150	7.62	0.166
15	11.45	0.328	11.81	0.313	11.16	0.309
Mean		0.148		0.132		0.121

Table 1 reports the design based and model based estimates. The design based estimates, $\hat{\theta}_{iT}$, is direct estimator based on design sampling (data of Susenas 2005, one year). The EBLUP estimates, $\hat{\theta}_{iT}^{P}$, use small are model with area effects (data of Susenas 2005, one year). Whereas, Hierarchical Bayes estimates, $\hat{\theta}_{iT}^{HB}$, use small are model with area and time effects (data of Susenas 2001 to 2005, five years). Estimated of mean square error denoted by $mse(\hat{\theta}_{iT})$, $mse(\hat{\theta}_{iT}^{P})$, and $mse(\hat{\theta}_{iT}^{HB})$. It is clear from Table 1 that estimated mean square error of model based is less than design based. The estimated mean square error of Hierarchical Bayes (time series data) is less than EBLUP (cross-sectional data).

6. Conclusion

Small area estimation can be used to increase the effective sample size and thus decrease the standard error. For such repeated surveys considerable gain in efficiency can be achieved by borrowing strength across both small area and time. Availability of good auxiliary data and determination of suitable linking models are crucial to the formation of indirect estimators.

7. Reference

- Datta, G.S., and Lahiri, P. (2000). A Unified Measure of Uncertainty of Estimated Best Linear Unbiased Predictors (BLUP) in Small Area Estimation Problems, *Statistica Sinica*, **10**, 613-627.
- 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.
- Pfeffermann, D. (2002). Small Area Estimation – New Developments and Directions. *International Statistical Review*, **70**, 125-143.
- Pfeffermann, D. and Tiller, R. (2006). State Space Modelling with Correlated Measurements with Aplication to Small Area Estimation Under Benchmark Constraints. S3RI Methodology Working Paper M03/11, University of Southampton. Available from: http://www.s3ri.soton.ac.uk/publications.
- Rao, J.N.K. (2003). Small Area Estimation. John Wiley & Sons, Inc. New Jersey.
- Rao, J.N.K., dan Yu, M. (1994). Small Area Estimation by Combining Time Series and Cross-Sectional Data. *Proceedings of the Section on Survey Research Method*. American Statistical Association.
- Swenson, B., dan Wretman., J.H. (1992). The Weighted Regression Technique for Estimating the Variance of Generalized Regression Estimator. *Biometrika*, **76**, 527-537.
- Thompson, M.E. (1997). Theory of Sample Surveys. London: Chapman and Hall.